Two-Sample Least Squares Projection

David Pacini*

University of Bristol

April 18, 2016

Abstract

This paper investigates the problem of making inference about the coefficients in the linear projection of an outcome variable y on covariates (x, z) when data are available from two independent random samples; the first sample contains information on only the variables (y, z), while the second sample contains information on only the covariates. In this context, the validity of existing inference procedures depends crucially on the assumptions imposed on the joint distribution of (y, z, x). This paper introduces a novel characterization of the identified set of the coefficients of interest when no assumption (except for the existence of second moments) on this joint distribution is imposed. One finding is that inference is necessarily nonstandard because the function characterizing the identified set is a nondifferentiable (yet directionally differentiable) function of the data. The paper then introduces an estimator and a confidence interval based on the directional differential of the function characterizing the identified set. Monte Carlo experiments explore the numerical performance of the proposed estimator and confidence interval.

KEYWORDS: Identification; Least Squares Projection; Data Combination.
JEL CLASSIFICATION: C21, C26

*Author address: Department of Economics, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK; Email: David.Pacini@bristol.ac.uk; Tel.: +44 (0) 11 79 28 84 37

1. Introduction

Least squares (or linear) projection coefficients are employed to approximate conditional expectations while guarding against misspecification and the curse of dimensionality (see Goldberger, 1991, or Hayashi, 2001, for a textbook exposition). Economists who use survey data for making inference about these coefficients often face the situation when no single sample includes all the variables of interest but there are two independent samples and each variable is included in at least one.[1] Complications arise because the coefficients of interest depend on moments of variables that are not jointly observed. The prominent method adopted to sidestep these complications is to impose additional assumptions on the distribution of the variables of interest. These include either restricting the dependence between the variables observed in different samples or requiring the presence of an instrumental variable observed in all samples (see e.g., the survey by Ridder and Moffitt, 2007). These assumptions are not testable. If there is doubt about their validity, then it is worth analyzing the sensitivity of inference to a failure of them. Little is known, however, about making inference when no assumption on the joint distribution of the variables of interest is imposed, except that the coefficients of interest are not point identified.

Motivated by the previous situation, we study the problem of making inference on the coefficients α and β in the linear projection $y = x'\alpha + z'\beta + u$ of an outcome variable y on covariates (z, x) when data are available from two independent random samples; the first sample gives information on only (y, z), while the second sample gives information on only the covariates. The disturbance term u is assumed to be uncorrelated with the covariates, and no assumption on the joint distribution of (y, x, z), except for the existence of second moments, is imposed. We show that the collection of values of the coefficients of interest compatible with knowledge of the distributions of (y, z) and of (z, x) (that is, the identified set) can be written as the intersection of two sets. We then derive a function characterizing the boundary of the identified set and use this function to construct an estimator and a confidence interval for the coefficients of interest.

The construction of the estimator and the confidence interval requires some elaboration be-

[1] Examples include Japelli, Pischke and Souleles (1998); Meghir and Palme (1999); Carroll, Dynan and Krane (2003); Fang, Keane and Silverman (2008); Bostic, Stuart and Painter (2009); and Brzozowski, Gervais, Klein, and Suzuki (2010). Additional examples are discussed in the text (see Section 2.1) and in the survey papers by Chesher and Nesheim (2006) and Ridder and Moffitt (2007).

cause the function characterizing the identified set involves the composition of max and min operations. It is well-known (see e.g., Hirano and Porter, 2012) that this type of operations render analog estimators systematically biased and invalidate the use of standard tools for inference (e.g., normal or nonparametric bootstrap approximation of sampling distributions). To overcome these difficulties, we construct an estimator which corrects the precision of the analog estimator of the identifed set. The correction is similar to the nonparametric bootstrap bias correction but it is based on a version of the bootstrap that is different from the nonparametric one. This version is introduced to overcome the inconsistency of the nonparametric bootstrap in our context. The confidence interval is constructed by inverting a test statistic. Both the estimator and the confidence interval are based on an approximation to the directional differential of the function characterizing the identified set. The theoretical properties of the estimator and the confidence interval are discussed. Monte Carlo experiments illustrate the implementation and evaluate the numerical properties of the proposed procedures.

1.1. Related Literature

The problem of making inference on least squares projection coefficients from two independent samples has been studied in several strands of literature under different concerns and methodologies. A first strand of literature focuses on matching-based estimation of linear regression coefficients (see e.g., Rassler, 2002; D'Orazio, DiZio and Scanu, 2006). Complications arising from the lack of observations on (y, x) are sidestepped by imputing the values of the covariates in the first sample (or the values of the outcome variable in the second sample; see e.g., Rodgers, 1984; Rubin, 1986; Moriarity and Scheuren, 2003). The imputation procedures are valid under the assumption that the outcome variable and the covariates observed on only one sample are independent conditional on the covariates observed on both samples. This conditional independence assumption often find little justification in practice. Our approach is thus useful to see what is lost when this conditional independence assumption is not valid.

A second strand of literature studies estimation and inference when instrumental variables are available (see e.g., Klevmarken, 1982; Angrist and Krueger, 1992: Arellano and Meghir, 1992; Inoue and Solon, 2010; Ichimura and Martinez-Sanchis, 2010). Complications arising from the lack of observations on (y, x) are overcome by assuming that some of the variables common to both samples are instrumental variables. Although this assumption does deliver

point identification, it is often the case that instrumental variables are not available. Our results are useful for this particular case.

The last strand of related literature focuses on nonparametric identification of the conditional expectation of the outcome variable given the covariates when the common covariates are discrete (e.g., Vitale, 1979; Cross and Manski, 2002; Molinari and Peski, 2006). In this strand of literature, identification analysis is carried out without imposing additional assumptions delivering point identification. Our work is in the same spirit, but it applies to a different setting. First, our focus is on identification and inference and not just identification. Second, we study least squares linear projections rather than conditional expectations. Third, we do not restrict the common variables to be discrete. Bontemps, Magnac and Maurin (2009) study identification of least squares projections from two independent samples. We consider the same setup but our characterization of the identified set is different. As we discuss below, our characterization does make use of the marginal distributions of (y, z, x) while their characterization does not.

To close this review of the literature, we mention the differences between our problem and other problems studied in the literature on sample combination. The assumption that the two samples are independent distinguishes our problem from the one with common observational units (e.g., Devereux and Triphati, 2009; Komarova, Nekipelov and Yakovlev, 2012; Poirer and Ziebarth, 2014). The fact that the two samples do not deliver point identification distinguishes this paper from either the case when the two samples jointly deliver point identification (see e.g., Chen, Hong and Tamer, 2005; Hirukawa and Prokhorov, 2014) or the case when one sample alone delivers point identification and a second sample is used for efficiency gains (see e.g., Hellerstein and Imbens, 1999). Fan, Sherman and Shum (2014) consider the related problem of combining samples to identify distributional treatment effects.

1.2. Organization of the Paper and Notation

The outline of the paper is as follows. In the next section, we define the coefficients of interest, describe the data, and discuss potential applications. In Section 3, we characterize the identified set and discuss what additional assumptions can shrink the identified set to a singleton. Section 4 introduces an estimator and a confidence interval for the coefficients of interest. In Section 5, we explore via Monte Carlo exercises the numerical performance of the

proposed estimator and confidence interval. Section 6 concludes. Three appendices collect the proofs.

We consider a collection of observational units (i.e., individuals, households, etc.) to be studied at a given period in time and index an observational unit in this collection by i. For each i, we define the random vector (y_i, x_i', z_i') on a probability space with probability measure P_o. We suppress the subscript i in the notation whenever this can be done without causing confusion. The outcome variable y is univariate and the covariates (x', z') are random vectors of dimension d_x and d_z, respectively. We use $\mathbb{E}$ to denote the expectation associated to P_o. We let $s^o_{xy} := \mathbb{E}(xy)$ denote the value of the expectation of the product of x and y. Similarly, we define $s^o_{xx'} := \mathbb{E}(xx')$, $s^o_{zz'} := \mathbb{E}(zz')$, etc.. We denote the d-dimensional Euclidean space by $\mathbb{R}^d$ and the unit sphere in $\mathbb{R}^d$ by $\mathbb{S}^d$.

2. The Setup

We begin by describing the parameter of interest.

Assumption P *(Parameter of Interest). Knowledge is sought about the coefficients* $\theta_o = (\alpha_o', \beta_o')'$ *defined by:*

(P.i) $$y = x'\alpha_o + z'\beta_o + u \quad \textit{with} \quad \mathbb{E}\big((x', z')'u\big) = 0,$$

where the joint distribution F^o_{yxz} *of* (y, x', z') *is such that:*

(P.ii) The variance of (y, x', z') *is finite and positive semidefinite.*

An equivalent way of writing (P.i) is $\theta_o := \arg\min_{(\alpha,\beta)} \mathbb{E}[\mathbb{E}(y|z, x) - x'\alpha - z'\beta)^2]$, which shows that θ_o can be interpreted as the coefficients in the least squares projection of the conditional expectation of y given (x', z') under quadratic loss. Since (P.i) does not restrict the conditional expectation of y given (x, z) to be a linear function, it is weaker than the mean-independence restriction $\mathbb{E}(y - x'\alpha_o - z'\beta_o|x, z) = 0$. The difference between (P.i) and the mean-independence restriction is often overlooked. In our context however, this distinction is of importance because lack of correlation and mean-independence deliver different identification results. (P.ii) ensures enough variation to define θ_o.

If a sample with replications of the triplet (y, x', z') was available, inference on θ_o would be straightforward. Here we focus on the case when replications of this triplet are unavailable. We assume instead that data are available from two samples:

Assumption D *(Data). Let $y, z \mapsto G^o_{yz}(y,z)$ denote the (y,z)-marginal distribution of $y, x, z \mapsto F^o_{yxz}(y,x,z)$. A similar notation is adopted for $x, z \mapsto G^o_{xz}(x,z)$. Data are available from two independent samples. The first sample $\{y_i, z_i'\}_{i=1}^{n_A}$ contains independent and identically distributed (iid) replications of (y, z') generated from G^o_{yz} for a group of n_A observational units. The second sample $\{x_j', z_j'\}_{j=n_A+1}^{n}$ contains iid replications of (x', z') generated from G^o_{xz} for a group of different $n_B = n - n_A$ observational units.*

2.1. Potential Applications

To illustrate the applicability of our setup, we now discuss potential applications fitting it. The first application comes from the work by Bostic, Gabriel, and Painter (2009, BGP from now on). They are interested in measuring α_o when y_i is the log consumption of household i living in the US in 2001, x_i is the log of housing wealth, z_i is a vector of household characteristics (including income and household size) and u_i is a disturbance term uncorrelated with z_i and x_i. Since a single sample of households with information on (y, x', z') is not available, BGP employ data from two samples; the Consumer Expenditure Survey (CEX) and the Survey of Consumer Finances (SCF). The CEX provides information on only households' consumption and demographic characteristics. The SCF in turn provides information on only demographic characteristics and housing wealth.

The second potential application concerns the measurement of returns to education. Let y_i denote log hourly wages for a worker i, z_i a vector of worker characteristics including education and experience, and let x_i denote a proxy for worker ability such as the intelligence quotient (IQ) test score. Interest is in the coefficient in β_o associated to education. Only a few datasets contain measures of wages, education, experience and IQ test scores for a single sample of workers. On one hand, household surveys carried out by government agencies usually gather information on wages, education and experience but not IQ test scores. On the other hand, there are surveys carried out by psychometricians gathering information on education and IQ test scores but not wages and experience.

The third potential application comes from the marketing literature. To design marketing campaigns, firms would like to infer the association, as measured by α_o, between the units y_i of a good purchased by consumer i and consumer's time exposure to advertising x_i. Collecting in-

formation on purchases and ads exposure for a single survey of consumers would be valuable but it is usually a very expensive proposition (see The Nielsen Company, 2009; Ipsos MORI, 2011). The common alternative is to have access to two independent samples. A first sample gathers information on purchases and consumers demographic characteristics z_i. A second sample contains information on ads exposure and the same consumers demographic characteristics.

3. Identification

In this section, we first define the *identified set* and describe the identification problem. We then introduce a characterization of this set. This characterization is the main result of the paper. We also discuss how additional assumptions can shrink the identified set to a singleton.

3.1. The Identification Problem

For identification purposes, assume that $y, z \mapsto G^o_{yz}(y,z)$ and $x, z \mapsto G^o_{xz}(x,z)$ are known. For a component x_k of x, let $x_k, y \mapsto F^o_{ky}(x_k, y)$ denote the joint distribution function of x_k and y. From (P.i), write the partitioned population normal equations as:

$$\begin{pmatrix} s^o_{xy} \\ s^o_{zy} \end{pmatrix} = \begin{pmatrix} s^o_{xx'} & s^o_{xz'} \\ s^o_{zx'} & s^o_{zz'} \end{pmatrix} \theta_o. \tag{1}$$

All the expectations in (1) are known except for s^o_{xy}. Solving the identification problem involves to exploit the restrictions imposed by Assumptions P and D on s^o_{xy} to ultimately recover θ_o.

To study how Assumptions P and D restrict θ_o, define $d_o := -[s^o_{zz'} - s^o_{zx'}(s^o_{xx'})^{-1}s^o_{xz'}]^{-1}s^o_{zx'}(s^o_{xx'})^{-1}$, $c_o := [s^o_{zz'} - s^o_{zx'}(s^o_{xx'})^{-1}s^o_{xz'}]^{-1}s^o_{zy}$, $b_o := (s^o_{xx'})^{-1} - (s^o_{xx'})^{-1}s^o_{xz'}d_o$ and $a_o := -(s^o_{xx'})^{-1}s^o_{xz'}c_o$. From (1), work out θ_o and write it as the composition "$\circ$" of two linear functions:

$$\theta_o := h \circ g(F^o_{1y}, .., F^o_{ky}, .., F^o_{d_x y}),$$

where the first linear function is

$$F_{1y}, .., F_{d_x y} \mapsto g(F_{1y}, .., F_{d_x y}) := \left(\int x_1 y dF_{1y}, .., \int x_{d_x} y dF_{d_x y} \right)'$$

and the second linear function is

$$\lambda \mapsto h(\lambda) := \begin{pmatrix} a_o + b_o\lambda \\ c_o + d_o\lambda \end{pmatrix}.$$

Let Λ denote the set of values of the expectation of the product between y and x such that the variance of (y, x', z') is positive semidefinite. Define the set Θ_M of values of θ_o compatible with this variance restriction by:

$$\Theta_M := \{\theta \in \mathbb{R}^{d_z+d_x} \text{ such that } \theta = h(\lambda) \text{ for any } \lambda \in \Lambda\}.$$

Let $\mathcal{F}$ denote the set of joint distributions for the pairs $(y, x_1), .., (y, x_{d_x})$ compatible with knowledge of $y, z \mapsto G^o_{yz}(y, z)$ and $x, z \mapsto G^o_{xz}(x, z)$. Define the set Θ_F of values of θ_o compatible with the marginal distributions by:

$$\Theta_F := \{\theta \in \mathbb{R}^{d_z+d_x} : \theta = h \circ g(F_{x_1y}, .., F_{x_ky}, .., F_{x_{d_x}y}) \text{ for any } (F_{x_1y}, .., F_{x_ky}, .., F_{x_{d_x}y}) \in \mathcal{F}\}.$$

Intersecting Θ_M and Θ_F one has:

Definition 1 *(Identified Set). The identified set Θ_I of θ_o delivered by Assumptions P and D is $\Theta_I := \Theta_M \cap \Theta_F$.*

The identification problem is to derive a characterization of Θ_I suggesting an estimation procedure. To our knowledge, a solution to this problem is not readily available in the literature. A partial solution has been proposed by Bontemps, Magnac and Maurin (2009), who characterize a set different from Θ_I. The set in Bontemps, Magnac and Maurin (2009) relates only to α_o and it is defined only by the restriction that the variance of (y, x', z') is positive semidefinite. By contrast, Θ_I is defined also by the restriction that the marginals of the joint distribution of (y, x', z') must be equal to the distributions characterizing the available data.

3.2. Solving the Identification Problem

To proceed, we obtain a first characterization of Θ_I in terms of *support functions*.[2] This characterization is attractive because it boils the identification problem down to solving two

[2]The support function of a convex set is equal to the signed distance of supporting hyperplanes of the set from the origin (see e.g., Hiriart-Urruty and Lemarechal, 2004).

optimization problems. The set Λ is convex because it is defined by a quadratic inequality (see Lemma 4 below). The set $\mathcal{F}$ is convex as well because it is defined by linear equalities (see Lemma 5 below). The sets Θ_M and Θ_F are linear transformations of the convex sets Λ and $\mathcal{F}$, respectively, because $\lambda \mapsto h(\lambda)$ and $F_{1y}, .., F_{d_x y} \mapsto h \circ g(F_{1y}, .., F_{d_x y})$ are linear. Since convexity is preserved under linear transformations (see Hiriart-Urruty and Lemarechal, 2004, Proposition 1.2.4), Θ_F and Θ_M are convex. Furthermore, since convex sets are characterized by their support functions (see Hiriart-Urruty and Lemarechal, 2004, Theorem 2.2.2), and the intersection of convex sets is equal to the minimum of their support functions (see Rockafellar, 1970, Corollary 16.5.1), Θ_I can be rewritten as follows.

Lemma 1 *(Characterization of the Identified Set). Let q denote a vector of directions belonging to the unit sphere $\mathbb{S}^{d_z+d_x}$ in $\mathbb{R}^{d_z+d_x}$. Then,*
(i) The set Θ_M is characterized by:

$$\Theta_M = \Big\{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq s_M(q) := \sup_{\lambda \in \Lambda} q'h(\lambda) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\Big\},$$

where $q \mapsto s_M(q)$ is the support function of Θ_M.
(ii) The set Θ_F is characterized by:

$$\Theta_F = \Big\{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq s_F(q) := \sup_{(F_{1y},..,F_{d_x y}) \in \mathcal{F}} q'h \circ g(F_{1y}, .., F_{d_x y}) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\Big\},$$

where $q \mapsto s_F(q)$ is the support function of Θ_F.
(iii) The identified set Θ_I is characterized by:

$$\Theta_I = \Big\{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq s_I(q) := \inf_{t \in \{M,F\}} s_t(q) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\Big\}, \tag{2}$$

where $q \mapsto s_I(q)$ is the support function of Θ_I.

Lemma 1 characterizes Θ_I as the collection of vectors θ whose linear combination with the vector of directions q is smaller or equal than the minimum of the value functions in $\sup_{\lambda \in \Lambda} q'h(\lambda)$ and $\sup_{(F_{1y},..,F_{d_x y}) \in \mathcal{F}} q'h \circ g(F_{1y}, .., F_{d_x y})$. We next study the solution to these optimization

problems, first for the case when x is univariate and then for the multivariate case.

3.3. The Univariate Case

Assuming that x is univariate simplifies the exposition. The restriction on the variance of (y, x', z') holds if and only if the determinant of this variance is nonnegative. In the next lemma, we use this observation to show that Λ is an interval. We further characterize the endpoints of this interval, and use this characterization to solve the optimization problem characterizing Θ_M.

Lemma 2 *(Operational Characterization of Θ_M when $d_x = 1$). Let q denote a vector belonging to the unit sphere in $\mathbb{R}^{d_x+d_z}$. Split q into $q = (q'_\alpha, q'_\beta)'$, where q_α is dimension d_x and q_β is of dimension d_z. Define $v_{oq} := q'_\alpha a_o + q'_\beta c_o$ and $e_{oq} := b'_o q_\alpha + d'_o q_\beta$. Let Assumptions (P) and (D) hold with $d_x = 1$. Let $\mathbb{V}(y)$ and $\mathbb{V}(x)$ denote the variance of y and x, respectively. Let define ρ^o_{zy} as the element-by-element correlation between z and y.*[3] *Define similarly ρ^o_{zx}. Define the moments:*

$$\lambda^o_{Ml} := \mathbb{E}(y)\mathbb{E}(x) + \big[\mathbb{V}(y)\mathbb{V}(x)\big]^{1/2}\Big[\rho^{o'}_{zy}\rho^o_{zx} - \sqrt{(1-\rho^{o'}_{zx}\rho^o_{zx})(1-\rho^{o'}_{zy}\rho^o_{zy})}\Big],$$
$$\lambda^o_{Mu} := \mathbb{E}(y)\mathbb{E}(x) + \big[\mathbb{V}(y)\mathbb{V}(x)\big]^{1/2}\Big[\rho^{o'}_{zy}\rho^{o'}_{zx} + \sqrt{(1-\rho^{o'}_{zx}\rho^o_{zx})(1-\rho^{o'}_{zy}\rho^o_{zy})}\Big].$$

Then, the support function characterizing Θ_M is:

$$s_M(q) = \max_{r\in\{l,u\}} v_{oq} + e_{oq}\lambda^o_{Mr}. \tag{3}$$

To characterize Θ_F, recall that $s_F(q) = \sup_{F_{xy}\in\mathcal{F}} q'h \circ g(F_{xy})$. This programming problem is a variant of the Kantorovich optimal transportation problem $\sup_{F_{xy}\in\mathcal{F}} g(F_{xy})$. The difference is in the transformation $q'h(\cdot)$ applied to the total cost function $g(F_{xy})$. By exploiting existing closed-form solutions for the Kantorovich optimal transportation problem (see e.g., Rachev and Ruschendorf, 1998), we obtain the following result:

Lemma 3 *(Operational Characterization of Θ_F when $d_x = 1$). Let Assumptions (P) and (D) hold with $d_x = 1$. Define v_{oq} and e_{oq} as in Lemma 2. Let $y, z \mapsto G^o_{y|z}(y|z)$ denote the conditional*

[3] For a random vector z and a random variable y, the k-th component of the vector of correlation coefficients ρ_{zy} is $[\mathbb{E}(z_k y) - \mathbb{E}(z_k)\mathbb{E}(y)]/(\mathbb{V}(z_k)\mathbb{V}(y))^{1/2}$.

distribution of y given z and let $\tau, z \mapsto Q^o_{x|z}(\tau|z)$ denote the conditional quantile function of x given z. Define the moments:

$$\lambda^o_{Fl} := \mathbb{E}\Big[yQ^o_{x|z}\big(1 - G^o_{y|z}(y|z)|z\big)\Big] \quad , \quad \lambda^o_{Fu} := \mathbb{E}\Big[yQ^o_{x|z}\big(G^o_{y|z}(y|z)|z\big)\Big],$$

where the expectation is with respect to G^o_{yz}. Then, the support function characterizing Θ_F is:

$$s_F(q) = \max_{r\in\{l,u\}} v_{oq} + e_{oq}\lambda^o_{Fr}. \tag{4}$$

By intersecting the support functions in Lemma 2 and Lemma 3 according to Lemma 1, one has the following result:

Proposition 1 *(Operational Characterization of the Identified Set when $d_x = 1$). Let Assumptions (P) and (D) hold with $d_x = 1$. Define v_{oq}, e_{oq}, λ^o_{tr} for $t \in \{M, F\}$ and $r \in \{l, u\}$ as in Lemmas 2 and 3. Then, the support function characterizing Θ_I is*

$$s_I(q) = \min_{t\in\{M,F\}} \max_{r\in\{l,u\}} v_{oq} + e_{oq}\lambda^o_{tr}. \tag{5}$$

3.4. The Multivariate Case

We now derive a characterization of the identified set when x may be multivariate. Extending the characterization of s_M to the multivariate case requires some elaboration because Λ is not longer an interval but an ellipsoid. The following Lemma uses the projection of y on z and of x on z to obtain this extension.

Lemma 4 *(Operational Characterization of Θ_M). Let Assumptions (P) and (D) hold. Define v_{oq} and e_{oq} as in Lemma 2 with the corresponding change in dimension to accommodate the case $d_x \geq 1$. Define the projection of y on z and the projection of x on z by $y = \delta_o' z + \Sigma^{1/2}_{Ao} w_A$ and $x = \Pi_o' z + \Sigma^{1/2}_{Bo} w_B$, respectively, where $\delta_o := (s^o_{zz'})^{-1} s^o_{zy}$, Σ_{Ao} is the variance of the residual $y - \delta_o' z$, w_A is a unit variance random variable, $\Pi_o := (s^o_{zz'})^{-1} s^o_{zx'}$, Σ_{Bo} is the variance of the residual $x - \Pi_o' z$ and w_B is a unit variance random vector of dimension $d_x \times 1$. Define further*

$C_o := \Sigma_{Ao}\Sigma_{Bo}$ *and* $B_o := \Pi_o' s^o_{zy} + s^o_{xz'}\delta_o - \Pi_o' s^o_{zz'}\delta_o$. *Then, the support function characterizing* Θ_M *is*

$$s_M(q) = v_{oq} + \left(e_{oq}' C_o e_{oq}\right)^{1/2} + e_{oq}' D_o. \tag{6}$$

We now turn our attention to s_F. The strategy exploited to characterize this function when x is univariate carries over the multivariate case. In particular, the linearity of the objective function $F_{1y}, .., F_{d_x y} \mapsto h \circ g(F_{1y}, .., F_{d_x y})$ allows us to extend Lemma 3 by applying the method of proof there to each of the components of $h \circ g(F_{1y}, .., F_{d_x y})$:

Lemma 5 *(Operational Characterization of* Θ_F*). Let Assumptions (P) and (D) hold. Define* v_{oq} *and* e_{oq} *as in Lemma 2 with the corresponding change in dimension to accommodate the case* $d_x \geq 1$. *Let* $e_{oq,k}$ *denote the k-th element of* e_{oq}. *Define* $y, z \mapsto G^o_{y|z}(y|z)$ *as in Lemma 3 and let* $\tau \mapsto Q^o_{k|z}(\tau|z)$ *denote the conditional quantile function of* x_k *given* z. *Define the moments:*

$$\lambda^o_{Fkl} := \mathbb{E}\big[yQ^o_{k|z}\big(1 - G^o_{y|z}(y|z)|z\big)\big] \quad ; \quad \lambda^o_{Fku} := \mathbb{E}\big[yQ^o_{k|z}\big(G^o_{y|z}(y|z)|z\big)\big].$$

Then, the support function characterizing Θ_F *is*

$$s_F(q) = \sum_{k=1}^{d_x} \max_{r\in\{l,u\}} v_{oq} + e_{oq,k}\lambda^o_{Fkr}. \tag{7}$$

By combining Lemmas 4 and 5, one obtains the main result of the paper:

Theorem 1 *(Operational Characterization of the Identified Set). Let Assumptions (P) and (D) hold. Then,* Θ_I *is characterized by the support function:*

$$s_I(q) = \min\left(v_{oq} + (e_{oq}' A_o e_{oq})^{1/2} + e_{oq}' D_o \ , \ \sum_{k=1}^{d_x} \max_{r\in\{l,u\}} v_{oq} + e_{oq,k}\lambda^o_{Fkr}\right), \tag{8}$$

where v_{oq}, e_{oq}, λ^o_{Fkl}, λ^o_{Fku}, C_o *and* D_o *are defined as in Lemmas 4 and 5.*

3.5. Obtaining Point Identification

We have emphasized that the maintained assumptions deliver set identification of θ_o. In a

particular application, the identified set may be too wide to provide the desired information about θ_o. We now review the force of additional assumptions to achieve point identification. For the sake of exposition, we focus on the case when x is univariate. There is one restriction whose implications are immediate. If z and x are uncorrelated, β_o is point-identified and α_o is set-identified. This result corresponds to the absence of omitted variable problem as explained in Goldbergerg (1991). It suggests that the two samples may be not very informative about α_o but informative about β_o when the correlation between the covariates is small. We explore this point in the Monte Carlo experiments below. If at least one of the elements in β_o is zero, then α_o is point-identified. This is equivalent to assume that one of the common covariates is an instrumental variable. To see why, fix z to be a scalar. When $\beta_o = 0$, it follows from the identifying mapping that $\alpha_o = s^o_{yz}/s^o_{xz}$. In such a case, α_o is point identified because s^o_{yz} and s^o_{xz} are. If Assumption (P.i) is replaced by the mean-independence restriction $\mathbb{E}[(y - x'\alpha_o - z'\beta_o)|x, z] = 0$, then θ_o is point identified. This is because the mean-independence restriction implies that any measurable function of z, such as z_k^2, is uncorrelated with the disturbance term u. In such a case, any of these functions can be used as an instrument to point identify θ_o. Finally, if y is conditionally independent of x given z, then θ_o is also point identified. Under this conditional independence assumption, $\mathbb{E}(yx)$ is equal to $\mathbb{E}[\mathbb{E}(y|z)\mathbb{E}(x|z)]$. Point identification follows after evaluating $\lambda \mapsto h(\lambda)$ at $\lambda = \mathbb{E}[\mathbb{E}(y|z)\mathbb{E}(x|z)]$.

4. Estimation and Inference

To reflect sampling variability, we now drop the assumption that the distributions of the two samples are known. We estimate these distributions and employ Theorem 1 to construct an estimator and a confidence interval for the components of θ_o.

4.1. Estimand

We begin by describing the object to be estimated. Motivated by the applications discussed in Section 2, we are interested in a component θ_{ko} of θ_o rather than in θ_o itself. We then estimate the one-dimensional projection of the identified set on the k-axis. Since the identified set is convex (see Lemma 1), this one-dimensional projection is an interval. The endpoints can

be characterized using the support function in Theorem 1. For a given direction q, define

$$\eta_q^o := \left(v_{oq}, (e_{oq}' C_o e_{oq})^{1/2} + e_{oq}' D_o, e_{oq,1}\lambda_{F1l}^o, e_{oq,1}\lambda_{F1u}^o, .., e_{oq,d_x}\lambda_{Fd_xl}^o, e_{oq,d_x}\lambda_{Fd_xu}^o\right)'.$$

Let q_l denote the k-negative canonical direction (i.e., q_l is a vector taking value -1 in position k and zero elsewhere) and let $q_u = -q_l$ denote the k-positive canonical direction. For any possible value η_{q_b} of $\eta_{q_b}^o$, use $s_I(q)$ in Theorem 1 to define the bounding functions

$$\begin{aligned} m_l(\eta_l) &:= -s_I(q_l) = -\min\left(v_{q_l} + (e_{q_l}' A e_{q_l})^{1/2} + e_{q_l}' B \ , \ \sum_{k=1}^{d_x} \max_{r\in\{l,u\}} v_{q_l} + e_{q_l,k}\lambda_{Fkr}\right), \\ m_u(\eta_u) &:= s_I(q_u) = \min\left(v_{q_u} + (e_{q_u}' A e_{q_u})^{1/2} + e_{q_u}' B \ , \ \sum_{k=1}^{d_x} \max_{r\in\{l,u\}} v_{q_u} + e_{q_u,k}\lambda_{Fkr}\right), \end{aligned}$$

where $\eta_b = \eta_{q_b}$ to save on notation. Since $q \mapsto s_I(q)$ gives the signed distance of supporting hyperplanes of the identified set from the origin (see footnote 2), one has that

$$[\theta_k] = [m_l(\eta_l^o), m_u(\eta_u^o)]$$

is the one-dimensional projection of the identified set on the k-axis.

4.2. Estimator

A natural idea to estimate $[\theta_k]$ would be to employ the sample analog principle. This approach however may systematically under(over)-estimate the true value of the endpoints of $[\theta_k]$. This is due to the presence of the min and max operators in the bounding functions defining the endpoints of $[\theta_k]$. With the aim of improving over the sample analog estimator, we introduce a bias-corrected estimator. This estimator has three steps. In the first step, we estimate $y, z \mapsto G_{y|z}^o(y|z)$ and $\tau, z \mapsto Q_{x|z}^o(\tau|z)$ by nonparametric methods. In the second step, we estimate v_{oq}, e_{oq}, λ_{Fr}^o for $r \in \{l, u\}$, D_o and C_o by their sample analogs. We denote these estimates by $\hat{v}_q$, $\hat{e}_q$,$\hat{\lambda}_{Fr}$, etc. In the third step, a bias-correction term is subtracted to the sample analog estimator of the endpoints

$$\widehat{[\theta_k]}_\kappa := \left[m_l(\hat{\eta}_l) - \kappa_l \hat{b}_l, m_u(\hat{\eta}_u) - \kappa_u \hat{b}_u\right],$$

where κ_l and κ_u are constants between zero and one, and $\hat{b}_l$ and $\hat{b}_u$ are estimates of the bias of the sample analog estimator (to be described below). The constants κ_l and κ_u are included to control the amount of bias-correction and to avoid highly variable estimates. When $\kappa_l = 0$ and $\kappa_u = 0$, no bias-adjustment is attempted and $\widehat{[\theta_k]}_\kappa$ is just the sample analog estimator. When $\kappa_l = 1$ and $\kappa_u = 1$, a full bias-adjustment is attempted.

We now describe $(\hat{b}_l, \hat{b}_u)$. If $\eta_b \mapsto m_b(\eta_b)$ was differentiable, we could use the nonparametric bootstrap to consistently estimate the bias and correct for it. $\eta_b \mapsto m_b(\eta_b)$ however is nondifferentiable. This is due to the composition of *min* and *max* functions. $\eta_b \mapsto m_b(\eta_b)$ is though directionally differentiable (see Appendix B). This allows us to approximate the bias of the sample analog estimator using the following algorithm:

Algorithm 1 *(Step-by-Step Calculation of Bias Correction). Let $\hat{\dot{m}}_b\big(\hat{\eta}_b, \cdot\big)$ denote a uniform consistent estimator of the the directional differential of $\eta_b \mapsto m_b(\eta_b)$ (see Appendix B).*
Step 1 - Draw S pairs of bootstrap samples, say $\big\{\big(\{y_{is}, z_{is}\}_{i=1}^{n_A}, \{x_{js}, z_{js}\}_{j=n_A+1}^{n_B}\big)$, for s=1,..,S$\big\}$, by resampling with replacement from the samples $\{y_i, z_i\}_{i=1}^{n_A}$ and $\{x_j, z_j\}_{j=n_A+1}^{n_B}$.
Step 2 - Let $\hat{\eta}_b$ denote the estimate of the vector of nuisance parameters η_b^o. Let $\hat{\eta}_{bs}^\star$ denote the estimate of the vector of nuisance parameters η_b^o computed from the bootstrapped samples s. For each sample s, calculate $\hat{\xi}_{bs}^\star := \hat{\dot{m}}_b\big(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}_{bs}^\star - \hat{\eta}_b)\big)$ for $b \in l, u$, where $\hat{\dot{m}}_b$ is a consistent estimator of $\dot{m}_b$.
Step 3 - The estimate of the bias of the sample analog estimator is

$$\hat{b}_b := \frac{1}{S}\sum_{s=1}^{S} n_A^{-1/2}\hat{\xi}_{bs}^\star.$$

The next Theorem establishes the consistency of $(\hat{b}_l, \hat{b}_u)$.

Theorem 2 *(Consistent Bias Estimation). Let Assumptions P and D hold. Let $\mathcal{Q}$ and $\mathcal{G}$ denote the parameter spaces for $\tau, z \mapsto Q_k^o(\tau|z)$, for all $k = 1, .., d_x$, and $y, z \mapsto G^o(y|z)$, respectively. Equip these spaces with norms $\|\cdot\|_\mathcal{Q}$ and $\|\cdot\|_\mathcal{G}$. Let assume that there are nonparametric estimators $\hat{Q}_k$ and $\hat{G}$ of $\tau, z \mapsto Q_{k|z}^o(\tau|z)$ and $y, z \mapsto G_{y|z}^o(y|z)$, respectively. Let further assume:*
(C.1) $G_{y|z}^o \in \mathcal{G}$; $\hat{G} \in \mathcal{G}$ with probability tending to one; and, for any number $0 \leq \delta \leq 1/4$, $n_A^\delta ||\hat{G} - G_{y|z}^o||_\mathcal{G} = o_{P_o}(1)$.
(C.2) For all $k = 1, .., d_x$, $Q_{k|z}^o \in \mathcal{Q}$; $\hat{Q}_k \in \mathcal{Q}$ with probability tending to one; and, for any

number $0 \leq \delta \leq 1/4$, $n_B^{\delta}||\hat{Q}_k - Q^o_{k|z}||_{\mathcal{Q}} = o_{P_o}(1)$.

(C.3) For all z *and* k, *the density* $x \mapsto g^o_{k|z}(x_k|z)$ *associated with* $x \mapsto G^o_{k|z}(x|z)$ *is bounded and bounded away from zero.*

(C.4) There are functions $y_i, z_i \mapsto \varphi_{kl}(y_i, z_i)$ *and* $y_i, z_i \mapsto \varphi_{ku}(y_i, z_i)$ *such that:*

$$\begin{aligned}\mathbb{E}\big(y_i[\hat{Q}_k(1-\hat{G}(y_i,z_i)|z_i) - Q^o_{k|z}(1-G^o_{y|z}(y_i,z_i)|z_i)]\big) &= n_A^{-1}\sum_{i=1}^{n_A}\varphi_{kl}(y_i,z_i) + o_P(n_A^{-1/2}),\\ \mathbb{E}\big(y_i[\hat{Q}_k(\hat{G}(y_i,z_i)|z_i) - Q^o_{k|z}(G^o_{y|z}(y_i,z_i)|z_i)]\big) &= n_A^{-1}\sum_{i=1}^{n_A}\varphi_{ku}(y_i,z_i) + o_P(n_A^{-1/2}),\end{aligned}$$

with $\mathbb{E}(\varphi_{kl}(y_i,z_i)) = \mathbb{E}(\varphi_{ku}(y_i,z_i)) = 0$, $\mathbb{E}(\varphi_{kl}(y_i,z_i)^2) \leq \infty$ *and* $\mathbb{E}(\varphi_{ku}(y_i,z_i)^2) \leq \infty$.

Let $(b_l, b_u) := \mathbb{E}[m_l(\hat{\eta}_l) - m_l(\eta_l^o), m_u(\hat{\eta}_u) - m_u(\eta_u^o)]$ *denote the bias of the sample analog estimator* $[m_l(\hat{\eta}_l), m_u(\hat{\eta}_u)]$. *Define* $(\hat{b}_l, \hat{b}_u)$ *as in Algorithm 1. Then,* $\|(\hat{b}_l, \hat{b}_u)' - (b_l, b_u)'\| = o_{P_o}(1)$.

Algorithm 1 approximates the bias of the sample analog estimator by the Monte Carlo mean of the bootstrap quantity $\hat{\dot{m}}_b\big(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}^{\star}_{bs} - \hat{\eta}_b)\big)$. The approximation $(\hat{b}_l, \hat{b}_u)$ is different from the nonparametric bootstrap, which is the Monte Carlo mean of the bootstrap quantity $n^{1/2}[m_b(\eta^{\star}_{bs}) - m_b(\hat{\eta}_b)]$. It is different as well from from the plug-in approximation $\dot{m}_b\big(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}^{\star}_{bs} - \hat{\eta}_b)\big)$. The nonparametric and plug-in approximation are inconsistent in the present context due to the nondifferentiability of the bounding functions (see Fang and Santos, 2014).

The nondifferentiability of $\eta_b \mapsto m_b(\eta_b)$ has two further implications for the evaluation of $\widehat{[\theta_k]}_{\kappa}$. First, impossibility results for nondifferentiable functions (see e.g., Hirano and Porter, 2012, Theorem 2) imply that potential reductions in bias may be offset by an increase in variance. This implication does not preclude modifying procedures to mitigate the imprecision problem, but suggests that one should assess carefully the properties of the modified procedure. In the next section, we evaluate this bias-variance trade-off via Monte Carlo exercises. Second, it implies that standard notions of asymptotic efficiency (i.e., variance bounds associated to minimum variance unbiased estimators) will not lead to useful comparisons between different estimators. Given this situation, we rely again on Monte Carlo exercises to evaluate $\widehat{[\theta_k]}_{\kappa}$.

4.3. Confidence Interval

Consider now the problem of inference. To communicate sampling variability, one may wish to construct a confidence interval for θ_{ko}. Fulfilling this wish is a delicate issue because

the nondifferentiability of the bounding functions precludes the constructions of confidence intervals based on asymptotically normal approximations, bootstrapping or subsampling the sample analog of the endpoints of $[\theta_k]$. To deal with the issues raised by nondifferentiability, we consider the confidence interval $C_n := \{\theta_k \in \mathbb{R} : T_n(\theta_k) \leq \hat{q}_{1-\tau}\}$, where

$$T_n(\theta_k) := max\left(n^{1/2}\big[m_l(\hat{\eta}_l) - \theta_k\big], 0\right)^2 + min\left(n^{1/2}\big[m_u(\hat{\eta}_u) - \theta_k\big], 0\right)^2$$

and $\hat{q}_{1-\tau}$ is a simulated critical value computed according to:

Algorithm 2. *(Step-by-Step Calculation of $\hat{q}_{1-\tau}$).*
Step 1. For $\hat{\xi}^\star_s$ calculated as in Algorithm 1, calculate $\hat{T}^*_{ns} := \max\left(\hat{\xi}^\star_{ls}, 0\right)^2 + \min\left(\hat{\xi}^\star_{us}, 0\right)^2$.
Step 2. Fix $\tau \in (0,1)$. Set $\hat{q}_{1-\tau}$ equal to the $1-\tau$ empirical quantile of $\{\hat{T}^*_{ns}\}_{s=1}^S$.

The following Theorem establishes the validity of C_n:

Theorem 3 *(Locally Uniform Asymptotic Confidence Interval). Let Assumptions P, D and C.1-C.4 hold. Let $\mathcal{H}$ denote the parameter space for $\eta_o = (\eta^o_l, \eta^o_u)$. Let K denote any compact set of $\mathcal{H}$ containing η_o. Consider the $n^{1/2}$-contingent cone at η_o with respect to K:*

$$\mathcal{K} := \{d \in \mathcal{H} : \exists d_n \text{ satisfying } \lim_{n\to\infty} \|d_n - d\|_{\mathcal{H}} = 0,\ \eta_o + n^{-1/2} d_n \in K\}$$

and $n^{-1/2}$-local perturbation neighborhood of P_o: $\mathcal{P}_o := \{P_{\eta_o + n^{-1/2}d} : d \in \mathcal{K}\}$. Then,

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_o, \theta_k \in [\theta_k]} P\big(\theta_k \in C_n\big) = 1 - \tau.$$

Theorem 3 guarantees that C_n asymptotically covers θ_{ko} with the pre-specified level $1-\tau$ for any value of the nuisance parameters associated with a probability function in $\mathcal{P}_o$ and any value of θ_{ko} in the one-dimensional projection $[\theta_k]$. This includes the case when θ_{ko} is either point-identified (i.e., $m_l(\eta^o_l) = m_u(\eta^o_u) = \theta_{ko}$ because x and z are uncorrelated), interval-identified (i.e., $m_l(\eta^o_l) < m_u(\eta^o_u)$) or η^o_b is near a point of nondifferentiability of the bounding functions.[4]

5. Monte Carlo Experiments

[4]The statistic $T_n(\theta_k)$ is of the type considered by Fan and Park (2014) in the context of nonparametric inference for counterfactual means

In this section, we employ simulated data to evaluate the performance of the procedures described in the previous section. The experiments show that: (i) When the correlation between z and x is low, the two samples may be informative about β_o; (ii) When inference does not take the restrictions on the marginal distributions into consideration, the estimated intervals do not take the best advantage of the data, resulting in estimates wider than necessary.

5.1. Design of Experiments

For computational simplicity, we let x_i to be univariate. For $\theta_o := (\alpha_o, \beta_{1o}, \beta_{2o}) = (1, 0, 1)$, we generate y_i according to $y_i = \beta_{1o} + x_i\alpha_o + z_i\beta_{2o} + u_i$, where u_i is a standard normal random variable independent of the covariates. The joint distribution of (x_i, z_i) is bivariate normal with mean zero and unit variance. The design variable is the correlation ρ_{xz} between z_i and x_i. In order to create two independent samples, we split the n draws of (y, x, z) into two samples of size n_A and n_B, respectively. In the first sample, we drop the realized values of x. In the second sample, we drop the realized values of y. We choose $n_A = n_B \in \{250, 500, 1000\}$. The number of replications is 250.[5]

5.2. Performance Measures for Post-Simulation Analysis

To evaluate the finite-sample properties of different interval estimators, we now describe a performance measure. To the best of our knowledge, there is no widely accepted loss function to evaluate interval estimators. Given this state of affairs, we decide to use as loss function the mean squared error uniformly integrated over an interval (MSEI). To describe this function, let $\tilde{\theta}_{ks}$, for $s \in \{l, u\}$, denote any estimator of the endpoints of $[\theta_k]$. For the interval estimator $[\tilde{\theta}_k] := [\tilde{\theta}_{kl}, \tilde{\theta}_{ku}]$, the MSEI is defined as:

$$MSEI([\tilde{\theta}_k]) := E\bigg(\int_{\tilde{\theta}_{kl}}^{\tilde{\theta}_{ku}} (\theta_k - \theta_{ko})^2 \frac{d\theta_k}{\tilde{\theta}_{ku} - \tilde{\theta}_{lu}} \bigg).$$

[5] We find increasing the number of replications computationally costly, especially for the largest sample size in consideration (i.e., $n_A = n_B = 1,000$). For the smallest sample size (i.e., $n_A = n_B = 250$), we also has produced results (available upon request) for 1,000 Monte Carlo replications. The qualitative conclusions obtained from 250 replications are not affected.

Magnac and Maurin (2008) show that $MSEI([\tilde{\theta}_k])$ has the following decomposition:

$$MSEI([\tilde{\theta}_k]) = \underbrace{(\bar{\theta}_k - \theta_{ko})^2}_{\text{Dec}} + \underbrace{\frac{1}{3}\left(\frac{\bar{\theta}_{ku} - \bar{\theta}_{kl}}{2}\right)^2}_{\text{AL}} + \underbrace{\frac{1}{3}E\left((\tilde{\theta}_{kl} - \theta_{ko})^2 + (\tilde{\theta}_{ku} - \theta_{ko})^2 + (\tilde{\theta}_{kl} - \theta_{ko})(\tilde{\theta}_{ku} - \theta_{ko})\right)}_{\text{ASE}},$$

where $\bar{\theta}_{ks} := E(\tilde{\theta}_{ks})$ for $s \in \{l, u\}$ and $\bar{\theta}_k := (\bar{\theta}_{kl} + \bar{\theta}_{ku})/2$. The first term (denoted Dec) can be interpreted as the square of the familiar bias term. The second term (AL) can be interpreted as the specific ambiguity due to set identification instead of point identification. The third term (ASE) can be interpreted as the usual variance term. As pointed out by Magnac and Maurin (2008), the latter decomposition is an adaptation of the usual decomposition of the mean squared error to the case when identification is partial.[6]

5.3. Comparison of Estimators

We calculate five estimators. The first estimator $\widehat{[\theta_k]}_M$ is the sample analog of the one-dimensional projection of the set implied by the positive semidefinite restriction on the variance of (y, x, z'). The second estimator $\widehat{[\theta_k]}_F$ is the sample analog of the one-dimensional projection of the set implied by the marginal restrictions on the joint distribution of (y, x', z'). The third estimator $\widehat{[\theta_k]}_I$ is the sample analog of the one-dimensional projection of the identified set. The fourth and fifth estimators are bias-corrected estimators. $\widehat{[\theta_k]}_{\kappa=.5}$ attempts a partial bias-correction by setting $\kappa_l = \kappa_u = .5$. $\widehat{[\theta_k]}_{\kappa=1}$ attempts a full bias-correction by setting $\kappa_l = \kappa_u = 1$. These estimators are obtained after replacing $y, z \mapsto G^o_{y|z}(y|z)$ and $\tau, z \mapsto Q^o_{x|z}(\tau|z)$ by series of cubic splines estimators. We choose the number of knots K_n and L_n according to the rule $K_n = L_n = \lfloor n_A^{1/3} \rfloor$. Implementing the bias-corrected estimator requires to choose a tuning parameter for estimating the directional derivative. In the simulations, we set this parameter to $log(n_A)$ (see our discussion in Appendix B).[7]

[6] An alternative to the MSEI is a loss function weighting coverage and length of the interval estimators. We are not aware, however, of the use of this type of loss functions in the context of set identifying models. For the sake of completeness, we report the Monte Carlo coverage and average length as well.

[7] All the experiments were carried out in the program R using the libraries "splines" (to generate cubic spline basis) and "quantreg" (to estimate $\tau, z \mapsto Q^o_{x|z}(\tau|z)$).

TABLE 1. *Monte Carlo Experiments: Comparison of Different Estimators.*

		Covariate z ($\beta_{2o} = 1$))					Covariate x ($\alpha_o = 1$)				
Obs.		M	F	I	$\kappa = .5$	$\kappa = 1$	M	F	I	$\kappa = .5$	$\kappa = 1$
250	Mean	[.911,1.49]	[.934,1.47]	[.935,1.47]	[.965,1.44]	[.996,1.41]	[-1.41,1.41]	[-1.29,1.30]	[-1.29,1.30]	[-1.27,1.28]	[-1.25,1.26]
	Dec	.042	.041	.041	.041	.042	1.00	.988	.990	.988	.987
	AL	.028	.024	.024	.019	.014	.669	.563	.562	.544	.526
	ASE	.018	.018	.018	.018	.018	.003	.005	.004	.005	.005
	RMSEI	.299	.291	.291	.281	.274	1.29	1.24	1.24	1.24	1.23
	Cove.	72%	68%	68%	60%	53%	100%	100%	100%	100%	99%
	Length	.587	.538	.538	.478	.418	2.83	2.60	2.59	2.55	2.51
500	Mean	[.925,1.49]	[.938,1.47]	[.938,1.47]	[.961,1.45]	[.984,1.43]	[-1.41,1.41]	[-1.35,1.35]	[-1.34,1.34]	[-1.33,1.33]	[-1.31,1.31]
	Dec	.043	.043	.043	.043	.043	1.00	1.00	.999	1.00	1.00
	AL	.026	.024	.024	.021	.016	.669	.608	.606	.592	.578
	ASE	.008	.008	.008	.008	.008	.001	.002	.002	.002	.002
	RMSEI	.281	.276	.276	.269	.263	1.29	1.26	1.26	1.26	1.25
	Cove.	74%	72%	72%	65%	54%	100%	100%	100%	100%	100%
	Length	.566	.541	.539	.495	.451	2.83	2.70	2.69	2.66	2.63
1000	Mean	[.921,1.49]	[.931,1.48]	[.931,1.48]	[.947,1.46]	[.963,1.35]	[-1.41,1.41]	[-1.36,1.37]	[-1.36,1.36]	[-1.35,1.35]	[-1.33,1.34]
	Dec	.043	.043	.043	.043	.043	1.00	.991	.992	.992	.992
	AL	.027	.025	.025	.022	.021	.671	.622	.622	.611	.600
	ASE	.004	.004	.004	.004	.004	.001	.001	.001	.001	.001
	RMSEI	.274	.271	.271	.265	.261	1.29	1.27	1.27	1.26	1.26
	Cove.	82%	80%	80%	75%	70%	100%	100%	100%	100%	100%
	Length	.576	.555	.554	.522	.491	2.83	2.73	2.73	2.70	2.69

This table presents different measures describing the finite sample performance of different estimators of the coefficients. All details about this experiments are in Section 5.1. We set the correlation between covariates equal to $\rho_{zx} = .2$ and the number of knots in the estimation of the conditional quantile and distributions functions according to $K_n = L_n = \lfloor n_B^3 \rfloor$. The label "Obs." indicates the number of observations in each sample. "M" is the sample analog estimator based on the positive definite restriction of the variance of (y, x, z). "F" is the sample analog estimator based on the restrictions on the marginal distributions of (y, x, z). "I" is the sample analog estimator based on both the positive definite restriction and the restrictions on the marginal distributions of (y, x, z). "$\kappa = .5$" is the bias-corrected estimator described in Section 4.2. with bias-adjustment term equal to one half. "$\kappa = 1$" is the bias-corrected estimator described in Section 4.2. with bias-adjustment term equal to one. 'Dec' stands for decentering of the mid-point of the interval, 'AL' is the adjusted length of the interval, 'ASE' is the variance of the estimators of the bounds, and 'RMSEI' is the square root of the mean squared error uniformly integrated defined in Section 5.2. The number of Monte Carlo replications is 250.

Table 1 compares the finite performance of these five estimators. We set ρ_{xz} to .2. For all sample sizes, $\widehat{[\theta_k]}_F$ improves upon $\widehat{[\theta_k]}_M$. This improvement can be attributed to the shorter AL of $\widehat{[\theta_k]}_F$ (see the row labeled 'AL' in the Table). This suggests that ignoring the marginal restrictions on the joint distribution of the variables of interest may give up information about the coefficients of interest. $\widehat{[\theta_k]}_I$ offers modest improvements upon $\widehat{[\theta_k]}_F$. These improvements come again from a reduction in AL. $\widehat{[\theta_k]}_{\kappa=.5}$ and $\widehat{[\theta_k]}_{\kappa=1}$ both improve upon $\widehat{[\theta_k]}_I$. Their Dec and ASE terms are similar to those of $\widehat{[\theta_k]}_I$, but their AL are smaller. $\widehat{[\theta_k]}_{\kappa=1}$ performs better than $\widehat{[\theta_k]}_{\kappa=.5}$. The estimators recover, on average, the sign β_{2o}. We may attribute this result to the low correlation between the covariates.

5.4. Comparison of Confidence Intervals

We now consider the estimation of confidence intervals. We implement the confidence interval in Algorithm 2 for a $1-\tau = .95$ nominal confidence level. We call this confidence interval *Delta*. We use different values of ρ_{xz} to evaluate the uniform properties of confidence intervals. In the experiment with $\rho_{xz} = 0$, the data generating process delivers point-identification of β_{2o}. In the other two experiments, the data generating process delivers only set-identification.

Implementing the Delta confidence interval is computationally intensive. It is worth then to explore the properties of computationally less intensive alternatives. We construct percentile nonparametric bootstrap confidence intervals (called *NonParametric*). The lower (upper) endpoint of this confidence interval is the .025(.975)-quantile of the nonparametric bootstrap distribution of the sample analog estimator of the lower(upper) endpoint of the one-dimensional projection of the identified set. We expect this confidence interval to perform worse than the Delta confidence interval. The number of bootstrap replications is 250.

Tables 2 presents the actual coverage probability and the average length of the Nonparametric and Delta confidence intervals. The percentile nonparametric bootstrap confidence interval is unnecessarely conservative, in particular, for the coefficient that may be point-identified (i.e., compare the average length reported in Table 2 for the Nonparametric and Delta Confidence intervals on β_2). For medium-large sample sizes (e.g., 500 - 1000 observations in each sample), the Delta confidence interval resolves this issue.

TABLE 2. *Monte Carlo Experiments: Performance of Confidence Intervals.*

			Covariate z ($\beta_2 = 1$)			Covariate x ($\alpha_1 = 1$)		
Obs.	Procedure		.4	.2	0	.4	.2	0
250	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.63	3.49	2.62	2.70	3.01	3.04
	Delta	Coverage	94%	96%	94%	100%	100%	100%
		Length	1.46	.833	.452	2.75	2.71	2.68
500	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.48	3.34	2.45	2.67	2.97	2.99
	Delta	Coverage	96%	95%	96%	100%	100%	100%
		Length	1.39	.742	.321	2.79	2.75	2.77
1000	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.38	3.25	2.32	2.62	2.94	2.96
	Delta	Coverage	99%	96%	95%	100%	100%	100%
		Length	1.36	.700	.224	2.81	2.79	2.78

This table presents different measures describing the finite sample performance of confidence intervals for the coefficients. The label "Obs." indicates the number of observations in each sample. The number of Monte Carlo and bootstrap replications is 250.

6. Summary and Conclusions

Applied researchers interested in making inference about least squares projection coefficients are often confronted to the situation when the relevant variables are measured in two or more independent samples, neither of which contains information on all the variables of interest. When no additional assumptions are invoked, the literature has shown that the coefficients of interest are not point-identified (see e.g., Ridder and Moffit, 2007). This paper characterizes the identified set for the coefficients of interest and introduces a bias-corrected estimator and a confidence interval. The proposed estimator and confidence interval exploit the fact that the function characterizing the identified set is directionally differentiable.

There are at least two topics which deserve further research. The first topic relates to the choice of the smoothing parameters for the estimators of the identified set. The second topic concerns theoretical comparison of alternative estimation procedures for intervals with nondifferentiable endpoints.

Acknowledgments. *I wish to thank Thierry Magnac. Stephane Bonhomme, Christian Bontemps, Andrew Chesher, Fabiana Gomez, Gregory Jolivet, Pascal Lavergne, Nour Meddahi, Jorge Ponce, Christoph Rothe, Senay Sokullu, Frank Windmeijer the Associate Editor, two anonymous referees and seminar participants at the Workshop on Set Identified Models in Toulouse '13, University of Bristol, the European Winter Meeting of the Econometric Society '11, Banco Central del Uruguay, the EC^2 '10 meeting, Universidad Carlos III/Madrid, the NESG '10 meeting, the ENTER Jamboree '10 meeting, and Toulouse School of Economics for useful comments and suggestions. All remaining errors are my responsability.*

Appendix A: Proofs of the Results in Section 3

Proof of Lemma 2. The support function of Θ_M is equal to:

$$s_M(q) \;:=\; \sup_{\lambda\in\Lambda} q'h(\lambda) = \sup_{\lambda\in\Lambda}\big(v_{oq} + e_{oq}\lambda\big) = v_{oq} + \sup_{\lambda\in\Lambda} e_{oq}\lambda.$$

Since the objective function in the programming problem in the latter display is linear, we have:

$$s_M(q) \;=\; v_{oq} + e_{oq}\mathbf{1}(e_{oq} > 0)\sup_{\lambda\in\Lambda}\lambda + e_{oq}\mathbf{1}(e_{oq} \le 0)\inf_{\lambda\in\Lambda}\lambda.$$

A solution to these programming problems must occur at the boundary of the feasible set Λ. We now characterize the boundary points of Λ.

For any random vector a and random variable b, let ρ_{ab} denote the correlation between the elements of a and b. Consider the determinant of the correlation of $(y, z', x)'$:

$$1 + \rho'_{zy}\rho_{zx}\rho_{yx} + \rho_{xy}\rho'_{zy}\rho_{zx} - \rho_{yx}\rho_{yx} - \rho'_{zx}\rho_{zx} - \rho_{zy}\rho_{zy}.$$

The variance of $(y, z', x)'$ is positive semidefinite if and only if the latter determinant is nonnegative. Viewed as a function of ρ_{yx}, one can rewrite this determinant and its sign restriction as the quadratic inequality:

$$A\rho_{xy}\rho_{xy} + B\rho_{xy} + C \ge 0,$$

where $A := -1$, $B := 2\rho'_{zy}\rho_{zx}$ and $C := 1 - \rho'_{zx}\rho_{zx} - \rho'_{zy}\rho_{zy}$. Since A is negative, the solution to this quadratic inequality is the interval defined by the two real roots of the quadratic equation $A\rho_{yx'}\rho_{xy} + B\rho_{yx'} + C = 0$:

$$\rho^-_{yx'} := \frac{-B + \sqrt{B - 4AC}}{2A}; \rho^+_{yx'} := \frac{-B - \sqrt{B - 4AC}}{2A}.$$

Replacing A, B and C by their definitions and rearranging terms,

$$\begin{aligned}
\rho^-_{yx} &= \rho'_{zy}\rho_{zx} + \sqrt{(1 - \rho'_{zx}\rho_{zx})(1 - \rho'_{zy}\rho_{zy})} \\
\rho^+_{yx} &= \rho'_{zy}\rho_{zx} - \sqrt{(1 - \rho'_{zx}\rho_{zx})(1 - \rho'_{zy}\rho_{zy})}
\end{aligned}$$

Under Assumption P.ii, $\rho^-_{yx'}$ and $\rho^+_{yx'}$ are finite. Plugging the ρ^-_{yx} ρ^+_{yx} in $\mathbb{E}(yx) = \mathbb{E}(y)\mathbb{E}(x) + [\mathbb{V}(y)\mathbb{V}(x)]^{1/2}\rho_{yx}$, one has that Λ is an interval with finite endpoints λ^o_{Ml} and λ^o_{Mu}. To conclude, plug these endpoints back in the support function and notice that $e_{oq}\mathbf{1}(e_{oq} \le 0)\lambda^o_{Ml} + e_{oq}\mathbf{1}(e_{oq} > 0)\lambda^o_{Mu} = \max(e_{oq}\lambda^o_{Ml}, e_{oq}\lambda^o_{Mu})$. ■

Before proving Lemma 3, we re-state, in a notation suitable for our purposes, an existing result characterizing bounds on the expectation of the product of two random variables with given marginals.

Lemma A.1. *(Explicit Solution for the Monge-Kantorovich Problem - Rachev and Ruschendorf, 1998, Theorem 3.1.2). Let F^o be a distribution function on $\mathbb{R}^2$ with marginals G^o_y and G^o_k and let (y, x_k) be distributed according to F^o. Let $\mathcal{F}_{y,k}$ denote the family of distribution functions on $\mathbb{R}^2$ with given marginals G^o_y and G^o_k. Suppose that there is a right continuous function*

$y, x_k \mapsto c(y, x_k)$ *satisfying the so-called Monge condition:*

$$c(\tilde{y}, \tilde{x}_k) - c(y, \tilde{x}_k) - c(\tilde{y}, x_k) + c(y, x_k) \geq 0$$

for $\tilde{x}_k \geq x_k$, $\tilde{y} \geq y$, *and that* $E\big(c(y, x_k)\big)$ *exists and is finite. Then,*

$$\inf_{F \in \mathcal{F}_{y,k}} \int c(y, x) dF(y, x) = \int c(y, x_k) d \max\{G^o_y(y) + G^o_k(x_k) - 1, 0\}$$

and

$$\sup_{F \in \mathcal{F}_{y,k}} \int c(y, x) dF(y, x) = \int c(y, x_k) d \min\{G^o_y(y), G^o_k(x_k)\}.$$

The value function in the optimization problems above correspond to the function $F \mapsto E\big(c(y, x_k)\big)$ evaluated at the Hoeffding-Frechet distributions. With Lemma A1 at hand, we proceed with the proof of Lemma 3 in the text.

Proof of Lemma 3. Using the notation in the Lemma, one can write the support function of Θ_F as:

$$s_F(q) = \sup_{F_{xy} \in \mathcal{F}} q' h \circ g(F_{xy}) = \sup_{F_{yx} \in \mathcal{F}} \big(v_{oq} + e_{oq} g(F_{yx})\big) = v_{oq} + \sup_{F_{yx} \in \mathcal{F}} e_{oq} g(F_{yx}),$$

where the third equality follows because v_{oq} does not depend on F_{yx}. Since e_{oq} does not depend on F^o_{yx},

$$s_F(q) = v_{oq} + e_{oq} \inf_{F_{yx} \in \mathcal{F}} g(F_{yx}) \mathbf{1}(e_{oq} \leq 0) + e_{oq} \sup_{F_{yx} \in \mathcal{F}} g(F_{yx}) \mathbf{1}(e_{oq} > 0).$$

The optimization problems in the latter display can be re-writen as:

$$\begin{aligned} \inf_{F_{yx} \in \mathcal{F}} g(F_{yx}) &= \inf_{F_{yx|z} \in \mathcal{F}_{y|z,x|z}} \int yx \mathrm{d}F_{y,x|z}(y, x|z) \mathrm{d}F_z(s) ds, \\ \sup_{F_{yx} \in \mathcal{F}} g(F_{yx}) &= \sup_{F_{yx|z} \in \mathcal{F}_{y|z,x|z}} \int yx \mathrm{d}F_{y,x|z}(y, x|z) \mathrm{d}F_z(s) ds. \end{aligned}$$

Since $y, x \mapsto yx$ in the objective function satisfies the Monge Condition, it follows from Lemma A.1. (Explicit Solution to the Monge-Kantorovich Problem) that the solution ocurrs at the conditional Hoeffding-Frechet distributions:[8]

$$G^l_{yx|z}(y, x|s) := \max\{0, G^o_{y|z}(y|s) + G^o_{x|z}(x|s) - 1\} \ , \ G^u_{yx|z}(y, x|s) := \min\{G^o_{y|z}(y|s), G^o_{x|z}(x|s)\}.$$

The result in the Theorem follows after evaluating $F_{yx|z} \mapsto \int yx \ dF_{yz|z}(y, z|s) dG^o_z(s)$ at $G^l_{yx|z}$

[8] $c(\tilde{y}, \tilde{x}_k) - c(y, \tilde{x}_k) - c(\tilde{y}, x_k) + c(y, x_k) = c\tilde{y}\tilde{x}_k - y\tilde{x}_k - \tilde{y}x_k + yx_k = (\tilde{y} - y)(\tilde{x}_k - x_k)$, which is non-negative whenever $\tilde{x} \geq x$ and $\tilde{y} \geq y$ as required by the Monge Condition.

and $G^u_{yx|z}$. Consider first the evaluation at $G^l_{yx|z}$:

$$\begin{aligned}\lambda^o_{Fl} &:= \int_{\mathcal{Z}}\int_{\mathcal{Y}\times\mathcal{X}} yx\mathrm{d}G^l_{yx|z}(y,x|s)\mathrm{d}G^o_z(s)\\ &= \int_{\mathcal{Z}}\int_{\mathcal{Y}\times\mathcal{X}} yx\mathrm{d}\max\{0, G^o_{y|z}(y|s)+G^o_{x|z}(x|s)-1\}\mathrm{d}G^o_z(s).\end{aligned}$$

Let $Q^o_{y|z}(\tau,z)$ and $Q^o_{x|z}(\upsilon,z)$ denote, respectively, the conditional τ-quantile of y given z and the conditional υ-quantile of x given z. By using the substitutions $y = Q^o_{y|z}(\tau,z)$ and $x = Q^o_{x|z}(\upsilon|z)$

$$\lambda^o_{Fl} = \int_{\mathcal{Z}}\int_{[0,1]\times[0,1]} Q^o_{y|z}(\tau,z)\times Q^o_{x|z}(\upsilon,z)\mathrm{d}\max\{0,\tau+\upsilon-1\}\mathrm{d}G^o_z(z).$$

Since $\mathrm{d}\max\{0,\tau+\upsilon-1\}$ is different from zero only at $\tau+\upsilon-1=0$, one has:

$$\lambda^o_{Fl} = \int_{\mathcal{Z}}\int_{[0,1]} Q^o_{y|z}(\tau,z)\times Q^o_{x|z}(1-\tau,z)\mathrm{d}\tau\mathrm{d}G^o_z(z).$$

By the change-of-variable $\tau = G^o_{y|z}(y|z)$:

$$\lambda^o_{Fl} = \int_{\mathcal{Z}}\int_{\mathcal{Y}} y\times Q^o_{x|z}(1-G^o_{y|z}(y|z),z)\mathrm{d}G^o_{yz}(y,z) = \mathbb{E}\big[y\cdot Q^o_{x|z}(1-G^o_{y|z}(y|z)|z)\big],$$

where the expectation is with respect to the joint distribution of (y,z). By a similar reasoning, the evaluation at $G^l_{yx|z}$ yields $\lambda^o_{Fu} = \mathbb{E}[y\cdot Q^o_{x|z}(G^o_{y|z}(y|z)|z)]$.

To conclude, plug the expression for λ^o_{Fl} and λ^o_{Fu} back in the expression of the support function and note that $e_{oq}\mathbf{1}(e_{oq}\leq 0)\lambda^o_{Fl} + e_{oq}\mathbf{1}(e_{oq}>0)\lambda^o_{Fu} = \max\{e_{oq}\lambda^o_{Fl}, e_{oq}\lambda^o_{Fu}\}$. ■

Proof of Proposition 1. This result follows after replacing $s_M(q)$ and $s_F(q)$ in Lemma 1, Equation (2), by their characterizations in Lemmas 2 and 3. ■

Proof of Lemma 4. The support function of Θ_M is:

$$s_M(q) = \sup_{\lambda\in\Lambda} q'h(\lambda) = v_{oq} + \sup_{\lambda\in\Lambda} e'_{oq}\lambda.$$

The aim is to find a closed form expression for the value function in the programming problem in the latter display.

We begin by characterizing the set Λ in a way that is suitable to our purpose. If the variance of (y,x',z') is positive definite so is the variance of (w_A,z',w'_B). We then write the variance of (w_A,z',w'_B) in terms of the unknown expectation $\lambda = E(yx)$. By construction z and w_A are uncorrelated (i.e., $\mathbb{E}(zw_A)=0$), as well as z and w_B (i.e., $E(zw_B)=0$). If z includes a constant, w_A and w_B have zero mean (i.e., $\mathbb{E}(w_A)=0$ and $\mathbb{E}(w_B)=0$). Hence, the covariance between w_A and w_B is

$$\begin{aligned}\mathbb{C}(w_A,w_B) &= \mathbb{E}(w_Aw_B)\\ &= \mathbb{E}\big[\Sigma^{-1/2}_{Ao}(y-z'\delta_o)\Sigma^{-1/2}_{Bo}(x-\Pi'_oz)\big]\\ &= (\Sigma_{Ao}\Sigma_{Bo})^{-1/2}[\lambda-\Pi'_oE(zy)-E(xz')\delta_o+\Pi'_oE(zz')\delta_o]\\ &:= C_o^{-1/2}(\lambda-D_o) := \tilde{\lambda},\end{aligned}$$

where the first equality follows because the residuals w_A and w_B have zero mean, the second equality follows after replacing w_A and w_B by their definitions, the fourth one after rearraging terms and the last line is a definition. The variance of (w_A, z', w_B') can be written as:

$$M := \begin{pmatrix} 1 & 0_{d_z} & \tilde{\lambda}' \\ 0_{d_z} & \mathbb{C}(z, z') & 0_{d_z \times d_x} \\ \tilde{\lambda} & 0_{d_x \times d_z} & I_{d_x} \end{pmatrix},$$

where 0_{d_z} is a d_z-dimensional column vector of zeros, $0_{d_x \times d_z}$ is a $d_x \times d_z$ matrix of zeros and I_{d_x} is an $d_x \times d_x$ identity matrix. Let Ω denote the variance of (z, w_B). If the matrix M is positive definite, so is the Schur complement S of Ω in M:

$$S := 1 - (0_{d_z}', \tilde{\lambda}')\Omega^{-1}(0_{d_z}', \tilde{\lambda}')' = 1 - \tilde{\lambda}'\tilde{\lambda} = 1 - \|\tilde{\lambda}\|^{1/2},$$

where the first equality follows from the fact that z and w_B are uncorrelated and w_B has variance one and the last equality from the definition of the Euclidian norm. Since S is a scalar, the positive definite condition is equivalent to $\|\tilde{\lambda}\| \leq 1$. Use this inequality to define the ball:

$$\tilde{\Lambda} := \{\tilde{\lambda} \in R^{d_x} : \|\tilde{\lambda}\| \leq 1\}.$$

$\tilde{\lambda}$ is at the boundary of $\tilde{\Lambda}$ when $||\tilde{\lambda}|| = 1$. Λ is an ellipsoid because it is linear transformation of $\tilde{\Lambda}$ ($\lambda = C_o^{1/2}\tilde{\lambda} + D_o$ by construction).

Use $\lambda = C_o^{1/2}\tilde{\lambda} + D_o$ to rewrite $s_M(q)$ in terms of $\tilde{\lambda}$ as:

$$s_M(q) = v_{oq} + \sup_{\lambda \in \Lambda} e_{oq}'\lambda = v_{oq} + \sup_{\tilde{\lambda} \in \tilde{\Lambda}} e_{oq}'(C_o^{1/2}\tilde{\lambda} + D_o) = v_{oq} + \sup_{\tilde{\lambda} \in \tilde{\Lambda}} e_{oq}'C_o^{1/2}\tilde{\lambda} + e_{oq}'D_o.$$

Since the objective function $\tilde{\lambda} \mapsto e_{oq}'C_o^{1/2}\tilde{\lambda}$ is linear, a solution to the programming problem has to ocurr at the boundary of $\tilde{\Lambda}$, that is, when $||\tilde{\lambda}|| = 1$. Hence, the programing problem is solved at $\tilde{\lambda}^* = (e_{oq}'C_o e_{oq})^{-1/2}C_o^{1/2}e_{oq}$. To conclude, replace $\tilde{\lambda} = \tilde{\lambda}^*$ in $e_{oq}'A^{1/2}\tilde{\lambda}$. ■

Proof of Lemma 5. Let $g_{ky}(F_{ky})$ denote the k-element of $g(F_{1y}, .., F_{ky}, .., F_{d_x y})$. Using the notation in the Proposition, we can write the support function as:

$$\begin{aligned} s_F(q) &= \sup_{F_{1y},..,F_{ky},..,F_{d_x y} \in \mathcal{F}} q'h \circ g(F_{1y}, .., F_{ky}, .., F_{d_x y}) \\ &= v_{oq} + \sup_{F_{1y},..,F_{ky},..,F_{d_x y} \in \mathcal{F}} \sum_{k=1}^{d_x} e_{oq,k} g_{ky}(F_{ky}) \\ &= v_{oq} + \sum_{k=1}^{d_x} \sup_{F_{ky} \in \mathcal{F}} e_{oq,k} g_{ky}(F_{ky}), \end{aligned}$$

where the third equality follows because $F_{1y}, .., F_{d_x y} \mapsto g(F_{1y}, .., F_{d_x y})$ is linear. The rest of the proof parallels that of Lemma 2 and is not repeated here. ■

Proof of Theorem 1. This result follows after replacing $s_M(q)$ and $s_F(q)$ in Lemma 1, Equation (2), by their characterizations in Lemmas 4 and 5. ■

Appendix B: Proofs of the Results in Section 4

We begin by calculating the directional differential of the bounding functions using the following chain rule.

Lemma B.1 *(Chain Rule for Hadamard Directional Differentiable Functions - Shapiro, 1990, Proposition 3.6). Let $h \mapsto \phi(h)$ be Hadamard directional differentiable at h_o, and let $\phi \mapsto \Phi(\phi)$ be Hadamard directional differentiable at $\phi_o := \phi(h_o)$. Let $\dot{\phi}(h_o, d)$ and $\dot{\Phi}(\phi_o, r)$ denote the Hadamard directional differential of $h \mapsto \phi(h)$ and $\phi \mapsto \Phi(\phi)$, respectively. Then, the composite mapping $h \mapsto g(h) := \Phi \circ \phi(h)$ is Hadamard directional differentiable at h_o and the chain rule $\dot{g}(h_o, d) = \dot{\Phi}\big(\phi_o, \dot{\phi}(h_o, d)\big)$ holds.*

To employ the chain rule, set $h_o = \eta_{lo}$ and $g(h) = m_l(\eta_l)$. Re-write $\eta_l \mapsto m_l(\eta_l)$ as the composition of two functions. The first function is

$$\eta_l \mapsto \phi(\eta_l) := \begin{pmatrix} \phi_M(\eta_{q_l}) \\ \phi_F(\eta_{q_l}) \end{pmatrix} := \begin{pmatrix} v_{q_l} + (e_{q_l} A^{-2} e_{q_l}')^{1/2} + e_{q_l} B \\ \sum_{k=1}^{d_x} \max_{r \in \{l,u\}} v_{q_l} + e_{q_l} \lambda_{Fkr} \end{pmatrix}.$$

For a given vector $\phi := (\phi_M, \phi_F) \in \mathbb{R}^2$, the second function is

$$\phi \mapsto \Phi(\phi) := -\min\{\phi_M, \phi_F\}.$$

With this notation at hand, $m_l(\eta_l) = \Phi \circ \phi(\eta_l)$. From Fang and Santos (2014), on can deduce that the Hadamard directional differential of $\phi \mapsto \Phi(\phi)$ at $\phi_o := (\phi_M^o, \phi_F^o)$ in the direction $r := (r_M, r_F)$ is:

$$\dot{\Phi}(\phi_o, r) := \begin{cases} -r_{t^*} & \text{if } \phi_M^o \neq \phi_F^o \\ -min\{r_M, r_F\} & \text{if } \phi_M^o = \phi_F^o \end{cases}$$

with $t^* := \arg\min_{t \in \{M,F\}} \phi_t^o$. For ι denoting a conformable vector of ones, the Hadamard directional differential of $\eta_l \mapsto \phi_M(\eta_l)$ at η_l^o in the direction d is:

$$\dot{\phi}_M(\eta_l^o, d) = \iota' d.$$

Using the chain rule above, the Hadamard directional differential of $\eta_l \mapsto \phi_F(\eta_l)$ at η_l^o in the direction d is:

$$\dot{\phi}_F(\eta_l^o, d) = \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d)$$

with

$$\dot{\phi}_{Fk}(\eta_l^o, d) := \begin{cases} d_1 + d_{ks_k^*} & \text{if } e_{q,k}\lambda_{Fkl} \neq e_{q,k}\lambda_{Fku} \\ d_1 + \max\big(d_{kl}, d_{ku}\big) & \text{if } e_{q,k}\lambda_{Fkl} = e_{q,k}\lambda_{Fku} \end{cases},$$

where d_1, d_{kl} and d_{ku} are the 1st, $2+k$-th and $2+k+1$-th elements in d, respectively, and $s_k^* := \arg\max_{t \in \{l,u\}} e_{q,k}\lambda_{Fkt}$. The chain rule in Lemma B.1 then implies that the Hadamard directional differential of $\eta_l \mapsto m_l(\eta_l)$ at η_l^o in the direction d is:

$$\dot{m}_l(\eta_l^o, d) = \dot{\Phi}\Big(\phi_o, r = \big(\dot{\phi}_M(\eta_l^o, d), \dot{\phi}_F(\eta_l^o, d)\big)\Big)$$

Using a similar reasoning, one can obtain the Hadamard directional differential of $\eta_u \mapsto m_u(\eta_u)$. We write a more explicit expression in the following Lemma:

Lemma B.2 *(Hadamard Directional Differential). The Hadamard directional differential of*

$\eta_l \mapsto m_l(\eta_l)$ *and of* $\eta_l \mapsto m_l(\eta_l)$ *are*

$$\dot{m}_l(\eta_l^o, d) := \begin{cases} -\iota' d & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) < 0 \\ -\sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d) & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) > 0 \\ -\min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d)\right) & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) = 0 \end{cases}$$

and

$$\dot{m}_u(\eta_u^o, d) := \begin{cases} \iota' d & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) < 0 \\ \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_u^o, d) & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) > 0 \\ \min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d)\right) & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) = 0 \end{cases},$$

respectively

For a positive sequence δ_n diverging to infinity and $\delta_n/\sqrt{n}$ converging to zero, we estimate the directional differentials by:

$$\hat{\dot{m}}_l(\hat{\eta}_l, d) := \begin{cases} -\iota' d & \text{if } s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) < -\delta_n \\ -\sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_l, d) & \text{if } s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) > \delta_n \\ -\min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_l, d)\right) & \text{if } -\delta_n < s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) < \delta_n \end{cases}$$

and

$$\hat{\dot{m}}_u(\hat{\eta}_u, d) := \begin{cases} \iota' d & \text{if } s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) < -\delta_n \\ \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_u, d) & \text{if } s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) > \delta_n \\ \min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_u, d)\right) & \text{if } -\delta_n < s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) < \delta_n \end{cases}.$$

The next proposition establishes the consistency of $\hat{\dot{m}}_l$ and $\hat{\dot{m}}_u$.

Lemma B.3 *(Consistent Estimator of the Directional Differential). Let Assumptions P and D hold. Let further assume that* $\delta_n \uparrow \infty$, $n^{-1/2}\delta_n \downarrow 0$, *and the conditions C.1-C.4 in Theorem 2 (Asymptotic Properties of the Nuisance Parameter Estimator) hold. Then,*

$$||\hat{\dot{m}}(\hat{\eta}_n, d) - \dot{m}(\eta_o, d)|| = o_{P_o}(1)$$

for any $d \in \mathbb{R}^{10}$.

Proof. Fix an arbitrary d. Under $\delta_n \uparrow \infty$ and $n^{-1/2}\delta_n \downarrow 0$, for any $\eta_{bn} \to \eta_b^o$

$$\hat{\dot{m}}_b(\eta_{bn}, d) - \dot{m}_b(\eta_o, d) = o(1).$$

Since $\eta_b^o \in \mathbb{R}$, η_b^o is Borel measurable and separable. Under C.1-C.4, $\hat{\eta}_b - \eta_b^o = o_{P_o}(1)$ (see Lemma C.3 below). Hence, the Extended Continous Mapping Theorem (see e.g., van der Vaart and Wellner, 1996, Theorem 1.11.1) implies

$$||\hat{\dot{m}}_b(\hat{\eta}_n, d) - \dot{m}_b(\eta_o, d)|| = o_{P_o}(1)$$

for any d. Since the result in the latter display holds jointly for $b \in \{l, u\}$, they can be combined to obtain Lemma B.3. ■

To establish theoretical properties of the bias-corrected estimator, we verify the conditions of a result in Fang and Santos (2014). For the sake of completeness, we begin by re-stating this result in a notation suitable for our purposes:

Lemma B.4 *(Consistency of the Delta Bootstrap - Fang and Santos, 2014, Theorem 3.3). Consider a function $f : \mathbb{D}_g \subseteq \mathbb{D} \mapsto \mathbb{E}$ describing a parameter of interest $f_o := f(h_o)$, where $\mathbb{D}_f$ denotes the domain of f and h_o a nuisance parameter. Let assume that:*
FS Condition 2.1.
(i) $\mathbb{D}$ and $\mathbb{E}$ are Banach spaces with norms $||\cdot||_{\mathbb{D}}$ and $||\cdot||_{\mathbb{E}}$, respectively.
(ii) $f : \mathbb{D}_f \subseteq \mathbb{D} \mapsto \mathbb{E}$ is Hadamard directional differentiable at h_o tangentially to the set $\mathbb{D}_o \in \mathbb{D}$, where $\mathbb{D}_o$ is a subset of $\mathbb{D}$.
FS Condition 2.2.
(i) $h_o \in \mathbb{D}_f$ and there is an estimator $\hat{h}_n$ of this nuisance parameter such that, for some sequence of positive numbers $r_n \uparrow \infty$, the random element $r_n(\hat{h}_n - h_o)$ converges in distribution to a random element $\mathbb{G}_o$ in $\mathbb{D}$.
(ii) $\mathbb{G}_o$ is tight and its support is included in $\mathbb{D}_o$.
FS Condition 2.3.
(i) The Hadamard directional derivative $\dot{f}(h_o, d)$ can be continuously extended to $\mathbb{D}$.
FS Condition 3.1. Let $\hat{h}_n^\star$ denote the bootstraped version of the estimator $\hat{h}_n$. Let $\{X_l\}_{l=1}^n$ denote the data and let $\{W_l\}_{l=1}^n$ denote random weights.
(i) $\hat{h}_n^\star : \{X_l, W_l\}_{l=1}^n \mapsto \mathbb{D}_\theta$ with $\{X_l\}_{l=1}^n$ independent of $\{W_l\}_{l=1}^n$.
(ii) $\hat{h}_n^\star$ satisfies

$$\sup_{b \in BL_1(\mathbb{D})} \left| E\left[b\big(r_n(\hat{h}_n^\star - \hat{h}_n)\big) | \{X_l\}_{l=1}^n \right] - E[f(\mathbb{G}_o)] \right| = o_P(1),$$

where $BL_1(\mathbb{D})$ is the set of Lipchitz functionals from $\mathbb{D}$ to $\mathbb{R}$ whose level and Lipchitz constant are bounded by one.
FS Condition 3.2.
(i) The sequence $r_n(\hat{h}_n^\star - \hat{h}_n)$ is asymptotically measurable.
(ii) $b\big(r_n(\hat{h}_n^\star - \hat{h}_n)\big)$ is a measurable function of $\{W_l\}_{l=1}^n$ outer almost surely in $\{X_l\}_{l=1}^n$ for any continuous and bounded function $b : \mathbb{D} \mapsto \mathbb{R}$.
FS Condition 3.3. There is an estimator $\hat{\dot{f}}$ of $\dot{f}$ such that for every compact set $K \subseteq \mathbb{D}_o$, $K^\delta := \{a \in \mathbb{D} : \inf_{b \in K} ||a - b||_{\mathbb{D}} < \delta\}$, and every $\epsilon > 0$:

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P_o\left(\sup_{d \in K^\delta} \left\| \hat{\dot{f}}(\hat{h}_n, d) - \dot{f}(h_o, d) \right\|_{\mathbb{E}} > \epsilon \right) = 0.$$

Then,

$$\sup_{b \in BL_1(\mathbb{E})} \left| E\left[b\Big(\hat{\dot{f}}\big(\hat{h}_n, r_n(\hat{h}_n^\star - \hat{h}_n)\big) \Big) \Big| \{X_l\}_{l=1}^n \right] - E\left[b\big(\dot{f}(h_o, \mathbb{G}_o) \big) \right] \right| = o_{P_o}(1).$$

To prove Theorem 2, we now verify the conditions of Lemma B.4.

Proof of Theorem 2. Our case corresponds to $h_o = \eta_b^o$ and $f(h) = m_b(\eta_b)$ with $\eta_b \in \mathbb{D} = \mathbb{R}^{2+2d_x}$, $\mathbb{E} = \mathbb{R}$, and $\mathbb{D}_g = \mathbb{K}^{2+2d_x}$ for some set compact set $\mathbb{K}^{2+2d_x} \subseteq \mathbb{R}^{2+2d_x}$.

To verify FS Condition 2.1(i), let equip $\mathbb{D}$ and $\mathbb{E}$ with the sup-norm. Then, $\mathbb{D}$ and $\mathbb{E}$ are

complete normed vector spaces (i.e., Banach spaces; see Rudin, 1986, Chapter 5.1). To verify FS Condition 2.1(ii), we refer to our discussion of Lemma B.2 (Hadamard Directional Differential).

To verify FS Condition 2.2, we first notice that, by construction, the nuisance parameter $h_o = \eta_b^o$ lives in some compact set. To verify FS Condition 2.2 (i), without loss of generality one can set $\mathbb{K}^{2+2d_x}$ equal to that compact set. The estimator is $\hat{h}_n = \hat{\eta}_b$. Under the conditions of Lemma C.3 (Asymptotic Properties of the Nuisance Parameter Estimator), we can set $r_n = n_A^{1/2}$ and $\mathbb{G}_o$ equal to a multivariate normal random vector. To verify FS Condition 2.2 (ii), it suffices to note that a multivariate normal random vector $\mathbb{G}_o$ is tight (because any random vector $\mathbb{G}_o$ is tight: for every constant $\epsilon > 0$ there exists a constant κ such that $P_o(\|\mathbb{G}_o\| > \kappa) < \epsilon$) and its support belongs to $\mathbb{D}_o$.

To verify FS Condition 2.3.(i), note that $\mathbb{D}_o$ corresponds to a cone in an Euclidean space. The cone $\mathbb{D}_o$ is a closed set in the topology induced by the sup norm $||\cdot||_{\mathbb{D}}$. Since $\eta_b \mapsto m_b(\eta_b)$ is Lipchitz continuous, $d \mapsto \dot{m}_b(\eta_l^o, d)$ is continuous (see e.g., Shapiro, 1990). Since $\mathbb{D}_o$ is closed, then the continuity of $d \mapsto \dot{m}_b(\eta_l^o, d)$ and Theorem 4.1 in Dugundji (1951) imply that $\dot{m}_b$ admits a continuous extension to $\mathbb{D}$.

To verify FS Condition 3.1 (i), we refer to our discussion at the end of Appendix C about the validity of the non-parametric bootstrap to approximate the sampling distribution of the estimator of the nuisance parameters $\hat{\eta}_b$. To verify FS Condition 3.1 (ii), it suffices to prove the consistency of the law of $n_A^{1/2}(\hat{\eta}_b^\star - \hat{\eta}_b)$ conditional on the data for the distribution of $\mathbb{G}_o$. This consistency result is established in Lemma C.4 (Consistency of the Nonparametric Bootstrap).

In our case FS Condition 3.2 is satisfied by construction because $\hat{h}_n = \hat{\eta}_b$ and $\hat{h}_n^\star = \hat{\eta}_b^\star$ correspond to empirical and bootstraped empirical processes, respectively.

To verify FS Condition 3.3.(i), notice that $\|\hat{\dot{m}}_b(\hat{\eta}_b, d_1) - \hat{\dot{m}}_b(\hat{\eta}_b, d_2)\| \leq \kappa \|d_1 - d_2\|$ for some constant $\kappa > 0$ and all $d_1, d_2 \in \mathbb{D}$ (because $d \mapsto \hat{\dot{m}}_b(\hat{\eta}_n, d)$ is Lipchitz continuous). Since $d \mapsto \dot{m}_b(\eta, d)$ is L-continuous, by a result in Fang and Santos (2014, Lemma A.6), showing that, for any $d \in \mathbb{D}_o$, $||\hat{\dot{m}}_b(\hat{\eta}_b, d) - \dot{m}_b(\eta_b^o, d)||_{\mathbb{E}} = o_{P_o}(1)$ suffices for establishing condition 3.3.(i). This latter condition is verified in Lemma B.3 (Consistent Estimator of the Directional Differential).

Hence, Lemma B.4 (Consistency of the Delta Bootstrap) implies that:

$$\begin{aligned} S^{-1}\sum_{s=1}^{S} \hat{\xi}_{bs}^\star &= E\Big(\hat{\dot{m}}_b\big(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}^\star - \hat{\eta}_n)\big)\Big) + o_{P^\star}(1) \\ &= E\Big(\dot{m}_b\big(\eta_b^o, n_A^{1/2}(\hat{\eta}^\star - \hat{\eta}_n)\big)\Big) + o_{P^\star}(1) + o_{P_o}(1) \\ &= n_A^{1/2} E\big(m_b(\hat{\eta}_b) - m_b(\eta_b^o)\big) + o_{P^\star}(1) + o_{P_o}(1) + o_{P_o}(1), \end{aligned}$$

where the second equality follows because $\hat{\dot{m}}_b$ is a consistent estimator of $\dot{m}_b$ (see Lemma B.3 Uniform Consistent Estimator of Directional Differential), and the last equality follows from the Delta Method Theorem in Fang and Santos (2014, Theorem 2.1). ■

To establish theoretical properties of the confidence interval C_n, we verify the conditions of a result in a companion paper (Pacini, 2016). For the sake of completeness, we begin by re-stating this result in a notation suitable for our purposes:

Lemma B.5 *(Locally Uniform Confidence Interval - Pacini, 2016, Theorem 1). Let $\eta_o := h(P_o)$ be an unknown nuisance parameter defined by a known bijective function $h : \mathcal{P} \mapsto \mathbb{H}$ taking values in a space $\mathbb{H} \subset \mathcal{H}$. The unknown parameter of interest $\theta_o \in \Theta \subset \mathbb{R}$ satisfies the inequalities*

$$m_l(\eta_o) =: \theta_l \leq \theta_o \leq \theta_u := m_u(\eta_o),$$

where $m_b : \mathbb{H} \mapsto \mathbb{R}$ *for* $b \in \{l, u\}$ *are known (up to* η_o*) bounding functions. Let the following assumptions hold:*
Pa Condition 2. Let $\mathcal{H}$ *a separable complete normed vector space. For each* $b \in \{l, u\}$*, for some Lipchitz constant* $L_b > 0$ *and any* η_1 *and* $\eta_2 \in \mathcal{H}$*,* $\eta \mapsto m_b(\eta)$ *is L-continuous:*

$$\|m_b(\eta_1) - m_b(\eta_2)\|_{\mathbb{R}} \leq L_b \|\eta_1 - \eta_2\|_{\mathcal{H}} \tag{Pa 2.i}$$

and Gateaux directional differentiable at η_o*:*

$$\dot{m}_b(\eta_o, d) := \lim_{t_n \downarrow 0} t_n^{-1}[m_b(\eta_o + t_n d) - m_b(\eta_o)] \tag{Pa 2.ii}$$

is finite for all $d \in \mathcal{H}$*.*
Pa Condition 3. There is a function $\eta, d \mapsto \hat{\dot{m}}_b(\eta, d)$ *(that may depend on n) satisfying:*
(i) If $\lim_{n\to\infty} \eta_n = \eta_o$ *with* $\eta_n \in \mathcal{H}$ *and* $\eta_o \in H$*, then*

$$\left|\hat{\dot{m}}_b(\eta_n, d) - \dot{m}_b(\eta_o, d)\right| = o(1) \qquad \text{for any } d \in \mathcal{H}; \tag{Pa 3.i}$$

(ii) $d \mapsto \hat{\dot{m}}_b(\eta, d)$ *is L-continuous for any* η*:*

$$\left|\hat{\dot{m}}_b(\eta, \tilde{d}) - \hat{\dot{m}}_b(\eta, d)\right| \leq L_{m_b} \|\tilde{d} - d\|_{\mathcal{H}} \qquad \text{for any } \tilde{d}, d \in \mathcal{H}, \eta \in H \tag{Pa 3.ii}$$

for some positive constant L_{m_b}*.*
Pa Condition 4. There is an estimator $\hat{\eta}_n : \{w_i\}_{i=1}^n \mapsto \mathcal{H}$ *of* η_o *satisfying:*

$$\hat{\eta}_n - \eta_o = o_{P_o}(1) \tag{Pa 4.i}$$

and, for $P_n \in \mathcal{P}_o$*,* $\eta_n = h(P_n)$ *and* Z_{η_o} *denoting a tight random element taking values in* $\mathcal{H}$*,*

$$Z_{P_n,n} := n^{1/2}(\hat{\eta}_n - \eta_n) \underset{P_n}{\rightsquigarrow} Z_{\eta_o}, \tag{Pa 4.ii}$$

where the distribution of Z_{η_o} *can depend on* η_o *but not on* η_n*.*
Pa Condition 5. There is an approximation $\hat{Z}_n^\star$ *of* Z_{η_o} *satisfying*

$$\sup_{b \in BL_{\mathcal{H}}} E\big[b\big(\hat{Z}_n^\star\big)|\{w_i\}_{i=1}^n\big] - E[b(Z_{\eta_o})] = o_{P_o}(1). \tag{Pa 5}$$

Consider the confidence interval C_n *constructed according to:*
Step 1. For a large S, simulate $1, .., s, .., S$ *realizations of* $\hat{Z}_n^\star$*. Denote a given realization by* $\hat{Z}_{ns}^\star$*.*
Step 2. Calculate $\hat{T}_{ns}^\star := \max\left(\hat{\dot{m}}_l(\hat{\eta}_n, \hat{Z}_{ns}^\star), 0\right)^2 + \min\left(\hat{\dot{m}}_u(\hat{\eta}_n, \hat{Z}_{ns}^\star), 0\right)^2$*.*
Step 3. Fix $\tau \in (0, 1)$*. Set* $\hat{q}_{1-\alpha}$ *equal to the* $1 - \alpha$ *empirical quantile of* $\{\hat{T}_{ns}^\star\}_{s=1}^S$*.*
Step 4. Create a grid in Θ*. Let* θ_c *denote a point in this grid.*
Step 5. Calculate $T_n(\theta_c)$ *for each* θ_c *in the grid. Accept* θ_c *if* $T_n(\theta_c) < \hat{q}_{1-\alpha}$ *otherwise discard* θ_c*. Take the smallest and largest accepted values as the endpoints of* C_n*.*
If
Pa Condition 6. The limiting distribution of the test statistic $T_n(\theta)$ *is continuous and strictly increasing at its* $1 - \alpha$ *quantile* $q_{1-\alpha}$*.*
Pa Condition 7. For some $(\eta_o, \theta_o) \in \mathcal{H}_n \times \Theta_I$*,* $\xi_o = (0, 0)$*.*
Then, C_n *satisfies*

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}_o, \theta \in [\theta_l, \theta_u]} P\Big(\theta \in C_n\Big) = 1 - \tau.$$

Proof of Theorem 3. For $sum_{k=1}^{d_x} := \sum_{k=1}^{d_x}$, define $g_{1l}(\eta_1, \eta_2) := sum_{k=1}^{d_x} max(\eta_{1k}, \eta_{2k})$, $r_{1l}(g_{1l}) := -g_{1l}$, $g_{2l}(r_{1l}, \eta_3) := max(r_{1l}, -\eta_3)$, $r_{2l}(g_{2l}) := -g_{2l}$. Rewrite the lower bounding function as

$$m_l(\eta) = -min\left(\sum_{k=1}^{d_x} max(\eta_{1k}, \eta_{2k}), \eta_3 \right) = r_{2l}\Big(g_{2l}\Big(r_{1l}\big(g_{1l}(\eta_1, \eta_2)\big), \eta_3 \Big) \Big).$$

We have already verified that η_o belongs to a Banach space (see Proof of Theorem 2). Since η_o is a point in an Euclidean space, η_o also belongs to a separable space. Since the max and sum functions are L-continuous, and L-continuity is preserved under composition of L-continuous functions, $\eta_1, \eta_2 \mapsto g_{1l}(\eta_1, \eta_2)$ and $r_{1l}, \eta_3 \mapsto g_{2l}(r_{1l}, \eta_3)$ are L-continuous (Pa 2.i). To verify Pa 2.ii, we refer to our discussion of Lemma B.2 (Hadamard Directional Differential). From the definition of $\hat{\dot{m}}_b$ above, notice that Pa 3.i and Pa 3.ii hold by construction. The nuisance parameters η_o are population moments and their sample analogs satisfy Pa 4.i and Pa 4.ii with Z_{η_o} a zero-mean multivariate normal vector with asymptotic variance depending on η_o (see Lemma C.3 Asymptotic Properties of Nuisance Parameter Estimator). The approximation $\hat{Z}_n^\star$ in Pa 5 can be constructed using the bootstrap (see Lemma C.4 Consistency of the Nonparametric Bootstrap). Since in this application $\hat{Z}_{\eta_o}$ is Gaussian (see Lemma C.3 Asymptotic Properties of Nuisance Parameter Estimator) and $\dot{m}_b(\eta_b^o)$ is non zero, Pa 3 and Pa 4 by a result in Davydov, Lifschitz and Smorodina (1998, Theorem 11.1) imply Pa 6. Pa 7 is satisfied when either the covariates observed in different samples are uncorrelated or there is an instrumental variable observed in both samples (see the discussion of point identification in Section 3). ■

Appendix C: Estimation of Nuisance Parameters

This appendix describes estimators for the nuisance parameters and establishes some of its asymptotic properties. These asymptotic properties in turn are useful to establish the properties of the inference procedures described in the text.

We begin by describing the estimators of components of the support functions. The estimators are:

$$\begin{aligned}
\hat{d} := -(\hat{s}_{zz'} - \hat{s}_{zx'}\hat{s}_{xx'}^{-1}\hat{s}_{xz'})^{-1}\hat{s}_{zx'}\hat{s}_{xx'}^{-1} \quad &; \quad \hat{c} := [\hat{s}_{zz'} - \hat{s}_{zx'}\hat{s}_{xx'}^{-1}\hat{s}_{xz'}]^{-1}\hat{s}_{zy} \\
\hat{b} := \hat{s}_{xx'}^{-1} - \hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{d} \quad &; \quad \hat{a} := -\hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{d} - \hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{c} \\
\hat{e}_q := q_\alpha'\hat{b} + q_\beta'\hat{d} \quad &; \quad \hat{v}_q := \hat{a}'q_\alpha + \hat{c}'q_\beta \\
\hat{A} := \hat{\Sigma}_A\hat{\Sigma}_B \quad &; \quad \hat{B} := \hat{\Pi}\hat{s}_{zy} + \hat{s}_{zx'}\hat{\delta} - \hat{\Pi}\hat{s}_{zz'}\hat{\delta} \\
\hat{\lambda}_{Fkl} := n_A^{-1}\sum_{i=1}^{n_A} y_i\hat{Q}_k\big(1 - \hat{G}(y_i|z_i)|z_i\big) \quad &; \quad \hat{\lambda}_{Fku} := n_A^{-1}\sum_{i=1}^{n_A} y_i\hat{Q}_k\big(\hat{G}(y_i|z_i)|z_i\big),
\end{aligned}$$

where $\hat{s}_{zz'} := n^{-1}\sum_{i=1}^n z_i z_i'$, $\hat{s}_{zx'} := n_B^{-1}\sum_{j=n_A+1}^n z_j x_j'$, $\hat{s}_{xx'} := n_B^{-1}\sum_{j=n_A+1}^n x_j x_j'$, $\hat{s}_{yz} := n_A^{-1}\sum_{i=1}^{n_A} y_i z_i$ and $y, z \mapsto \hat{G}(y|z)$ and $\tau, z \mapsto \hat{Q}_k(\tau|z)$ are non-parametric estimators for the conditional distribution $y, z \mapsto G_{y|z}^o(y|z)$ and the conditional quantile function $\tau, z \mapsto Q_{k|z}^o(\tau|z)$, respectively.

Let define $\hat{\eta} := (\hat{\eta}_{q_l}', \hat{\eta}_{q_u}')'$ with

$$\begin{aligned}
\hat{\eta}_{q_l} &:= (\hat{v}_{q_l}, (\hat{e}_{q_l}'\hat{A}\hat{e}_{q_l})^{1/2} + \hat{e}_{q_l}'\hat{B}, \hat{e}_{\underline{q},1}\hat{\lambda}_{F1l}, \hat{e}_{q_l,1}\hat{\lambda}_{F1u}, .., \hat{e}_{q_l,d_x}\hat{\lambda}_{Fd_xl}, \hat{e}_{q_l,d_x}\hat{\lambda}_{Fd_xu})', \\
\hat{\eta}_{q_u} &:= (\hat{v}_{q_u}, (\hat{e}_{q_u}'\hat{A}\hat{e}_{q_u})^{1/2} + \hat{e}_{q_u}'\hat{B}, \hat{e}_{q_u,1}\hat{\lambda}_{F1l}, \hat{e}_{q_u}\hat{\lambda}_{F1u}, .., \hat{e}_{q_u,1}\hat{\lambda}_{Fd_xl}, \hat{e}_{q_u,1}\hat{\lambda}_{Fd_xu})'.
\end{aligned}$$

The components of $\hat{\eta}$ are either sample analogs of unconditional moments (e.g., $\hat{v}_{ql}$) or sample analogs of unconditional moments with unknown functions estimated non-parametrically (e.g., $\hat{e}_{ql,k}\hat{\lambda}_{Fkl}$). If there were no unknown functions estimated nonparametrically, convergence in probability for $\hat{\eta}$ and convergence in distribution for $(\hat{\eta}-\eta_o)\sqrt{n_A}$ could be established using a Law of Large Numbers and a Central Limit Theorem for sums of i.i.d. random variables. The presence of unknown functions estimated nonparametrically makes the direct application of these standard results an inviable approach. A viable approach is to use the Law of Large Numbers and the Central Limit Theorem for semiparametric estimation problems involving both finite and infinite dimensional unknown parameters as in Chen, Linton and Van Keilegom (2003). We next re-state these results in a notation suitable for our purposes.

Lemma C.1 *(Law of Large Numbers - Chen, Linton and Van Keilegom, 2003, Theorem 1). Assume that data $\{X_l\}_{l=1}^n$ is randomly sampled from a distribution P whose support is a proper subset of $\mathbb{R}^d$. Let denote $\mathcal{A}$ for a finite dimensional parameter set and Ψ for an infinite dimensional parameter set. Let equip these sets with norms $||\cdot||_{\mathcal{A}}$ and $||\cdot||_{\Psi}$, respectively. For some positive integers p and k, assume that there exists a measurable vector-valued function $X_l,\gamma,\psi \mapsto m(X_l,\gamma,\psi):\mathbb{R}^d\times\mathbb{R}^p\times\Psi\mapsto\mathbb{R}^k$, and define the nonrandom function $\gamma,\psi\mapsto M(\gamma,\psi):=E(m(X_l,\gamma,\psi)):\mathbb{R}^p\times\Psi\mapsto\mathbb{R}^k$. Let denote $\gamma_o\in\mathcal{A}$ and $\psi_o\in\Psi$ as the true unknown finite and infinite dimensional parameters. Define the sample analog $M_n(\gamma,\psi):=n^{-1}\sum_{l=1}^n m(X_l,\gamma,\psi)$ and assume there is a nonparametric estimator $\hat{\psi}$ of ψ_o. Define the estimator $\hat{\gamma}:=\arg\min_\gamma ||M_n(\gamma,\hat{\psi})||$. Suppose further that:*
CLK Condition (1.1). $||M_n(\hat{\gamma},\hat{\psi})||\leq\inf_{\gamma\in\mathcal{A}}||M_n(\gamma,\hat{\psi})||+o_P(1)$.
CLK Condition (1.2). For all positive constant $\delta>0$, there exists another positive constant $\epsilon(\delta)>0$ such that $\inf_{||\gamma-\gamma_o||_{\mathcal{A}}>\delta}||M(\gamma,\psi_o)||\geq\epsilon(\delta)>0$.
CLK Condition (1.3). Uniformly for all $\gamma\in\mathcal{A}$, $\psi\mapsto M(\gamma,\psi)$ is continuous (with respect to the metric $||\cdot||_\Psi$) at $\psi=\psi_o$.
CLK Condition (1.4). $||\hat{\psi}-\psi_o||_\Psi=o_P(1)$.
CLK Condition (1.5'). For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n=o(1)$,

$$\sup_{\gamma\in\mathcal{A},||\psi-\psi_o||_\Psi<\delta_n}||M_n(\gamma,\psi)-M(\gamma,\psi)||=o_P(1)$$

Then, $\hat{\gamma}-\gamma_o=o_P(1)$.

Lemma C.2 *(Central Limit Theorem - Chen, Linton and Van Keilegom, 2003, Theorem 2). Assume that the conditions of the Law of Large Numbers in Lemma C.1 are satisfied. For some sequence $\{\delta_n\}$ of positive numbers converging to zero, define the shrinking sets $\mathcal{A}_{\delta_n}:=\{\gamma\in\mathcal{A}:\|\gamma-\gamma_o\|_{\mathcal{A}}\leq\delta_n\}$ and $\Psi_{\delta_n}:=\{\psi\in\Psi:\|\psi-\psi_o\|_\Psi\leq\delta_n\}$. For any (γ,ψ), let denote the derivative of $\gamma\mapsto M(\gamma,\psi)$ evaluated at γ by $\Gamma_1(\gamma,\psi)$ and the pathwise derivative of $\psi\mapsto M(\gamma,\psi)$ at ψ in the direction $(\bar{\psi}-\psi)$ by $\Gamma_2(\gamma,\psi)[\bar{\psi}-\psi]$. Assume further that:*
CLK Condition (2.1). $||M_n(\hat{\gamma},\hat{\psi})||\leq\inf_{\gamma\in\mathcal{A}}||M_n(\gamma,\hat{\psi})||+o_P(n^{-1/2})$.
CLK Condition (2.2). (i) The derivative $\gamma\mapsto\Gamma_1(\gamma,\psi_o)$ exists for $\gamma\in\mathcal{A}_{\delta_n}$ and is continuous at $\gamma=\gamma_o$;
(ii) The matrix $\Gamma_1:=\Gamma_1(\gamma_o,\psi_o)$ is of full (column) rank.
CLK Condition (2.3). For all $\gamma\in\mathcal{A}_{\delta_n}$, the pathwise derivative $\Gamma_2(\gamma,\psi_o)[\psi-\psi_o]$ of $M(\gamma,\psi_o)$ exists in all directions $[\psi-\psi_o]\in\Psi$; and for all $(\gamma,\psi)\in\mathcal{A}_{\delta_n}\times\Psi_{\delta_n}$ with a sequence δ_n of positive numbers converging to zero:
(i) $||M(\gamma,\psi)-M(\gamma,\psi_o)-\Gamma_2(\gamma,\psi_o)[\psi-\psi_o]||\leq\kappa||\psi-\psi_o||_\Psi^2$ *for some positive constant $\kappa>0$;*
(ii) $||\Gamma_2(\gamma,\psi_o)[\psi-\psi_o]-\Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o]||\leq o(1)\delta_n$.

CLK Condition 2.4. $\hat{\psi} \in \Psi$ *with probability tending one; and* $||\hat{\psi} - \psi_o||_\Psi = o_P(n^{-1/4})$.
CLK Condition 2.5'. For all sequences $\{\delta_n\}$ *of positive numbers converging to zero*

$$\sup_{||\gamma-\gamma_o||_{\mathcal{A}}<\delta_n, ||\psi-\psi_o||_\Psi<\delta_n} ||M_n(\gamma,\psi) - M(\gamma,\psi) - M_n(\gamma_o,\psi_o)|| = o_P(n^{-1/2}).$$

CLK Condition 2.6'. (i) $M_n(\gamma_o,\psi_o) = n^{-1}\sum_{i=1}^n m(X_l) + o_P(n^{-1/2})$ *with* $E[m(X_l)] = 0$ *and* $E[\|m(X_l)\|^2] < \infty$*; (ii)* $\Gamma_2(\gamma_o,\psi_o)[\hat{\psi}-\psi_o] = n^{-1}\sum_{i=1}^n \phi(X_l) + o_P(n^{-1/2})$ *with* $E[\phi(X_l)] = 0$ *and* $E[\|\phi(X_l)\|^2] < \infty$.
Then, $\sqrt{n}\big(\hat{\gamma} - \gamma_o\big)$ *converges in distribution to a random vector with multivariate normal distribution.*

We next verify the conditions of these two Lemmas to obtain the following properties for $\hat{\eta}$:

Lemma C.3 *(Asymptotic Properties of the Nuisance Parameter Estimator* $\hat{\eta}$*). Let Assumptions P and D hold. Set the infinite dimensional parameter space* Ψ *equal to the product of the spaces of functions* ψ_k *mapping the support of* (y,z) *into the support of* (x_k,x_k) *according to:*

$$\psi_k(y,z) := \begin{pmatrix} \psi_{lk}(y,z) \\ \psi_{uk}(y,z) \end{pmatrix} := \begin{pmatrix} Q_k\big(1 - G(y|z)|z\big) \\ Q_k\big(G(y|z)|z\big) \end{pmatrix},$$

where $\tau, z \mapsto Q_k(\tau,z)$ *is any conditional quantile function for the random variable* x_k *conditional on* z *and* $y, z \mapsto G(y|z)$ *is any conditional distribution function for the random variable* y *conditional on* z*. Denote by* $\mathcal{Q}$ *and* $\mathcal{G}$ *the parameter spaces for* $\tau, z \mapsto Q_k(\tau|z)$*, for all* $k = 1,..,d_x$*, and* $y, z \mapsto G(y|z)$*, respectively. Equip these spaces with norms* $\|\cdot\|_{\mathcal{Q}}$ *and* $\|\cdot\|_{\mathcal{G}}$*. Let further assume that C.1-C.4 hold.*
Then, $\hat{\eta} - \eta_o = o_{P_o}(1)$ *and* $n_A^{1/2}(\hat{\eta} - \eta_o)$ *converges in distribution to a random vector with a multivariate normal distribution.*

Proof. To verify the conditions for the Law of Large Numbers in Lemma C.1 and the Central Limit Theorem in Lemma C.2, we interpret the data $\{X_l\}_{l=1}^n$ as the two independent samples $\{y_i, z_i\}_{i=1}^{n_A}$ and $\{z_j, x_j\}_{j=n_A+1}^n$ described in Assumption D, and we assume that there is some number $0 < \kappa < 1$ such that, for $n_B = n - n_A$, as $n_A, n_B \to \infty$, $n_A/n \to \kappa$ and $n_B/n \to 1-\kappa$. Set γ_o to

$$\gamma_o = (\mu_y^o, \mu_x^o, \mu_z^o, s_{yy}^o, s_{xx'}^o, s_{zz'}^o, s_{yz}^o, s_{xz}^o, e_{\underline{q},1}^o \lambda_{F1l}^o, .., e_{\overline{q},d_x}^o \lambda_{Fd_xu}^o),$$

where $\mu_y = E(y_i)$ and similarly for μ_x and μ_z. Set the finite dimensional parameter space $\mathcal{A}$ equal to some compact subset $\mathbb{K}$ of the Euclidean space $\mathbb{R}^p$. In our case, the function $X_l, \gamma, \psi \mapsto m(X_l,\gamma,\psi) : \mathbb{R}^{1+d_x+d_z} \times \mathbb{R}^p \times \Psi \mapsto \mathbb{R}^p$ correspond to $\big(m_A(y_i,z_i,\gamma,\psi), m_B(z_j,x_j,\gamma), m_{AB}(z_l,\gamma), m_C(y_i,z_i,\gamma,\psi)\big)'$ with:

$$m_A(y_i,z_i,\gamma) := \begin{pmatrix} \mu_y - y_i \\ s_{yy} - y_i^2 \\ s_{zy} - z_i y_i \end{pmatrix},$$

$$m_B(z_j,x_j,\gamma) := \begin{pmatrix} \mu_x - x_j \\ vec(s_{xx'}) - vec(x_j x_j') \\ vec(s_{xz'}) - vec(x_j z_j') \end{pmatrix},$$

$$m_{AB}(z_l,\gamma) := \begin{pmatrix} \mu_z - z_l \\ vec(s_{zz'}) - vec(z_l z_l') \end{pmatrix},$$

$$m_C(y_i,z_i,\gamma,\psi) := \begin{pmatrix} e_{q_l,1}\lambda_{F1l} - (q_{l\alpha}b + q_{l\beta}d)y_i Q_1(1-G(y|z)|z) \\ e_{q_u,1}\lambda_{F1l} - (q_{u\alpha}b + q_{u\beta}d)y_i Q_1(G(y|z)|z) \\ \vdots \\ e_{q_l,d_x}\lambda_{Fd_xu} - (q_{l\alpha}b + q_{l\beta}d)y_i Q_{d_x}(1-G(y|z)|z) \\ e_{q_u,d_x}\lambda_{Fd_xu} - (q_{u\alpha}b + q_{u\beta}d)y_i Q_{d_x}(G(y|z)|z) \end{pmatrix}.$$

Set $M_n(\gamma,\psi)$ equal to

$$M_n(\gamma,\psi) := \begin{pmatrix} n_A^{-1}\sum_{i=1}^{n_A} m_A(y_i,z_i,\gamma) \\ n_B^{-1}\sum_{j=n_A+1}^{n} m_B(z_j,x_j,\gamma) \\ n^{-1}\sum_{l=1}^{n} m_{AB}(z_l,\gamma) \\ n_A^{-1}\sum_{i=1}^{n_A} m_C(y_i,z_i,\gamma,\psi) \end{pmatrix}.$$

The estimator $\hat{\gamma} := \arg\min_\gamma ||M_n(\gamma,\hat{\psi})||$ in our case corresponds to:

$$\hat{\gamma} = (\hat{\mu}_y, \hat{\mu}_x, \hat{\mu}_z, \hat{s}_{yy}, \hat{s}_{xx'}, \hat{s}_{zz'}, \hat{s}_{yz}, \hat{s}_{xz}, \hat{e}_{\underline{q},1}\hat{\lambda}_{F1l}, .., \hat{e}_{\overline{q},d_x}\hat{\lambda}_{Fd_xu}).$$

To verify CLK Condition (1.1), it suffices to note that in our case $||M_n(\hat{\gamma},\hat{\psi})||$ and $\inf_\gamma ||M_n(\gamma,\hat{\psi})||$ are both equal to zero because there are no over-identifying restrictions (i.e., $p=k$). To verify CLK Condition (1.2), notice that $M(\gamma,\psi_o)$ delivers point identification of the finite dimensional parameter γ_o by assumption. To verify CLK Condition (1.3), notice that $\psi \mapsto M(\gamma,\psi)$ in our case is a linear bounded operator for all γ, and then continuous at $\psi=\psi_o$. We now verify CLK Condition (1.4). Because we have assumed that $\hat{G}$ converges in probability to $G^o_{y|z}$ uniformly over y,z (see C.1) and that $\hat{Q}_k$ converges in probability to $Q^o_{k|z}$ uniformly over τ,z (see C.2), it follows that $\hat{\psi}_k(y,z) = \left(\hat{Q}_k\big(1-\hat{G}(y|z)|z\big), \hat{Q}_k\big(\hat{G}(y|z)|z\big)\right)$ converges in probability to $\psi_{ko}(y,z) = \left(Q^o_{k|z}\big(1-G^o_{y|z}(y|z)|z\big), Q^o_{k|z}\big(G^o_{y|z}(y|z)|z\big)\right)$ uniformly over (y,z). To verify CLK Condition (1.5'), we note that this condition is implied by CLK Condition (2.5'), which is verified below. Hence, by the Law of Large Numbers in Lemma C.1., we have that $\hat{\gamma}-\gamma_o = o_P(1)$. Moreover, since η_o is a continuous function of γ_o, we can apply the Continuous Mapping Theorem (see e.g., van der Vaart, 1998, Theorem 18.11) to conclude that $\hat{\eta}-\eta_o = o_P(1)$.

We now verify CLK Conditions (2.1) - (2.6'). To verify CLK Condition (2.1), notice that in our case $||M_n(\hat{\gamma},\hat{\psi})|| = \inf_\gamma ||M_n(\gamma,\hat{\psi})|| = 0$ because there are no over-identifying restrictions. Verifying CLK Condition (2.2) is standard, so we omit it here. To verify CLK Condition (2.3), start by noticing that $\psi \mapsto M(\gamma,\psi)$ is twice Frechet differentiable because it is linear. The

pathwise derivative is the first Frechet differential:

$$\Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o] = \begin{pmatrix} -eoq_l, 1E\Big(y[Q_1(1-G(y|z)|z) - Q^o_{1|z}(1-G^o_{y|z}(y|z)|z)]\Big) \\ \vdots \\ -e_{oq_u,1}E\Big(y[Q_{d_z}(G(y|z)|z) - Q^o_{d_z|z}(G^o_{y|z}(y|z)|z)]\Big) \end{pmatrix}$$

and the remainder $Rem(\psi-\psi_o) = M(\gamma,\psi) - M(\gamma,\psi_o) - \Gamma_2(\gamma,\psi_o)[\psi-\psi_o]$ satisfies

$$\|Rem(\psi-\psi_o)\| = O(\|\psi-\psi_o\|^2).$$

Hence, CLK Condition (2.3)(i) is satisfied. In our case, the difference $\Gamma_2(\gamma,\psi_o)[\psi-\psi_o] - \Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o]$ is linear in $\gamma-\gamma_o$. This implies that, for any γ in the shrinking set $\mathcal{A}_{\delta_n}$, the sequence $\|\Gamma_2(\gamma,\psi_o)[\psi-\psi_o] - \Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o]\|$ converges to zero as required by CLK Condition (2.3)(ii).

To verify CLK Condition (2.4), start from:

$$n_A^{1/4}\|\hat\psi-\psi_o\|_\Psi = n_A^{1/4}\sum_{k=1}^{d_x}\left\|\big(\hat Q_k(1-\hat G)\big),\hat Q_k(\hat G)\big) - \big(Q^o_{k|z}(1-G^o_{y|z})\big), Q^o_{h|z}(G^o_{y|z})\big)\right\|_{\Psi_k}.$$

It then suffices to verify the condition for any given k. Add-and-subtract $\big(Q^o_{k|z}(1-\hat G)\big), Q^o_{k|z}(\hat G)\big)$ to the left hand side and apply the Triangle Inequality

$$\begin{aligned} n_A^{1/4}\|\hat\psi_k-\psi_{ko}\|_{\Psi_k} &\le n_A^{1/4}\left\|\big(\hat Q_k(1-\hat G)\big),\hat Q_k(\hat G)\big) - \big(Q^o_{k|z}(1-\hat G)\big), Q^o_{k|z}(\hat G)\big)\right\|_{\Psi_k} \\ &+ n_A^{1/4}\left\|\big(Q^o_{k|z}(1-\hat G)\big), Q^o_{k|z}(\hat G)\big) - \big(Q^o_{k|z}(1-G^o_{y|z})\big), Q^o_{k|z}(G^o_{y|z})\big)\right\|_{\Psi_k}. \end{aligned}$$

C.2 implies that the first term in the right-hand-side of the latter display is $o_P(1)$. Consider now the second term in the right-hand-side. For some $\tilde G \in \mathcal{G}_{\delta_n}$, use the Inverse Function Theorem to obtain the following mean value expansion:

$$n_A^{1/4}\|\hat\psi_k-\psi_{ko}\|_{\Psi_k} \le o_P(1) + n_A^{1/4}\left\|\left(\frac{[G^o_{y|z}-\hat G]}{g^o_{k|z}\big(Q^o_{x|z}(1-\tilde G)\big)}, \frac{[\hat G - G^o_{y|z}]}{g^o_{k|z}\big(Q^o_{k|z}(\tilde G)\big)}\right)\right\|_{\Psi_k}.$$

Applying the Cauchy-Schwarz Inequality:

$$\begin{aligned} n_A^{1/4}\|\hat\psi_k-\psi_{ko}\|_{\Psi_k} &\le o_P(1) \\ &+ \left\|\Big(\big(g^o_{k|z}\big(Q^o_{k|z}(1-\tilde G)\big)\big)^{-1}, \big(g^o_{k|z}\big(Q^o_{k|z}(\tilde G)\big)\big)^{-1}\Big)\right\|_{\Psi_k} \\ &\times n_A^{1/4}\big\|\big(G^o_{y|z}-\hat G\big), \big(\hat G - G^o_{y|z}\big)\big\|_{\Psi_k}. \end{aligned}$$

Under the assumption that the density $x \mapsto g^o_{k|z}(x|z)$ is bounded away from zero and bounded for all z (see C.3), we have that $\left\|\Big(\big(g^o_{k|z}\big(Q^o_{k|z}(1-\tilde G)\big)\big)^{-1}, \big(g^o_{k|z}\big(Q^o_{k|z}(\tilde G)\big)\big)^{-1}\Big)\right\|_{\Psi_k}$ is bounded and

$$n_A^{1/4}\|\hat\psi_k-\psi_{ko}\|_\Psi \le o_P(1) + O(1)n_A^{1/4}\big\|\big(G^o_{y|z}-\hat G\big), \big(\hat G - G^o_{y|z}\big)\big\|_{\mathcal{G}\times\mathcal{G}}$$

The symmetry property of a norm and the convergence in probability in Assumption $C.1$ on the estimator of the distribution function implies that $n_A^{1/4}\left\|\left(G^o_{y|z}-\hat{G}\right),\left(\hat{G}-G^o_{y|z}\right)\right\|_{\mathcal{G}\times\mathcal{G}} = n_A^{1/4}\left\|\left(\hat{G}-G^o_{y|z}\right)\right\|_{\mathcal{G}} = o_P(1)$. Hence,

$$n_A^{1/4}\|\hat{\psi}_k-\psi_{ko}\|_\Psi \quad\leq\quad o_P(1)+O(1)o_P(1)\leq o_P(1).$$

To verify CLK Condition 2.5', notice that $\gamma\mapsto m_A(y_i,z_i,\gamma)$, $\gamma\mapsto m_B(z_j,x_j,\gamma)$ and $\gamma\mapsto m_{AB}(z_l,\gamma)$ are Holder continuous, and $\gamma,\psi\mapsto m_C(y_i,z_i,\gamma,\psi)$ is uniformly Lipchitz and twice continuously differentiable (because it is linear). Hence, a result in Chen, Linton and van Keilegom (2003, Theorem 3), implies that CLK Condition 2.5' (and 1.5') are satisfied.

Verifying CLK Condition (2.6')(i) is standard, so we omit it here. To verify CLK Condition (2.6')(ii), start from $\Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o]$. Using the asymptotic linear representation in C.4,

$$\Gamma_2(\gamma_o,\psi_o)[\psi-\psi_o]=\begin{pmatrix} n_A^{-1}\sum_{i=1}^{n_A}(-e_{oq_l})\varphi_{1l}(y_i,z_i)\\ \vdots \\ n_A^{-1}\sum_{i=1}^{n_A}(-e_{oq_u})\varphi_{d_zu}(y_i,z_i)\end{pmatrix}+o_P(n^{-1/2}).$$

Hence, to verify CLK Condition (2.6')(ii) it suffices to set

$$\phi(X_l)=\left((-e_{oq_l})\varphi_{1l}(y_i,z_i),..,(-e_{oq_u})\varphi_{d_zu}(y_i,z_i)\right)'.$$

Since $n_A^{1/2}(\hat{\eta}-\eta_o)$ is a differentiable function of $n_A^{1/2}(\hat{\gamma}-\gamma_o)$, we conclude from the Delta Method that $n_A^{1/2}(\hat{\eta}-\eta_o)$ converges in distribution to some random vector $\mathbb{G}_o$ with multivariate normal distribution. ■

We next provide conditions under which the non-parametric bootstrap can consistently estimate the asymptotic distribution of $(\hat{\eta}-\eta_o)n_A^{1/2}$. To obtain such conditions, we rely again on a Theorem by Chen, Linton and Van Keilegom (2003, Theorem B), which is re-stated below for convenience:

Lemma C.4 *(Consistency of the Non-Parametric Bootstrap - Chen, Linton, Van Keilegom, 2003, Theorem B). Let $\hat{\psi}^\star$ be the same estimator as $\hat{\psi}$ but based on bootstrap data. Here, and subsequently, superscript $\star$ denotes a moment computed under the bootstrap distribution conditional on the original data. Define the bootstrap estimator $\hat{\gamma}^\star := \|M_n^*(\hat{\theta}^\star,\hat{\psi}^\star)-M_n(\hat{\gamma},\hat{\psi})\|$. Suppose that $\{X_l\}_{l=1}^n$ is i.i.d.; $\gamma\in int(\mathcal{A})$; that CLK Conditions (2.1), (2.4), (2.5') and (2.6) hold with 'in probability' replaced by 'almost sure'; that CLK Conditions (2.2) holds with ψ_o replaced by $\psi\in\Psi_{\delta_n}$ while CLK Condition (2.3) holds with ψ_o replaced by $\psi\in\Psi_{\delta_n}$; and that $\gamma,\psi\mapsto\Gamma_1(\gamma,\psi)$ is continuous in ψ at $\gamma=\gamma_o,\psi=\psi_o$. Suppose:*
(2.4B) With $P^\star$-probability tending to one, $\hat{\psi}^\star\in\Psi$, and $\|\hat{\psi}^\star-\hat{\psi}\|_\Psi = o_{P^\star}(n^{-1/4})$.
(2.5'B) $\sup_{(\gamma,\psi)\in\mathcal{A}_{\delta_n}\times\Psi_{\delta_n}}\|M_n^\star(\gamma,\psi)-M_n(\gamma,\psi)+M_n^\star(\gamma_o,\psi_o)-M_n(\gamma_o,\psi_o)\| = o_{P^\star}(n^{-1/2})$ for all positive values $\delta_n=o(1)$.
(2.6B) $\sqrt{n}\|M_n^\star(\hat{\gamma},\hat{\psi})-M_n(\hat{\gamma},\hat{\psi})+\Gamma_2(\hat{\gamma},\hat{\psi})[\hat{\psi}^\star-\hat{\psi}]\|$ converges in distribution to a random vector with multivariate normal distribution.
Then, $(\hat{\gamma}-\gamma_o)\sqrt{n}$ converges in probability to $(\hat{\gamma}^\star-\hat{\gamma})\sqrt{n}$ in $P^\star$-probability.

CLK Conditions (2.4B)-(2.6'B) can be verified under the same assumptions implying CLK Conditions (2.4) and (2.6'), by using the corresponding asymptotic linear approximation for $\hat{\psi}^\star-$

$\hat{\psi}$. This observation establishes the consistency of the non-parametric bootstrap to approximate the distribution of the random vector $\mathbb{G}_o$. Following the results in Chen, Linton and Van Keilegom (2003, Theorem 2), we could derive a closed-form expression for the variance of $\mathbb{G}_o$. Such an expression however is unnecessary to prove the validity of the inference procedures in the text.

References

ANGRIST, J. and A. KRUEGER (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples", *Journal of the American Statistical Association* 87, pp. 328-36.

ARELLANO, M. and C. MEGHIR (1992): "Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets", *Review of Economic Studies*, 59(3) pp. 537-59.

BONTEMPS, C., T. MAGNAC and E. MAURIN (2009): "Set Identified Linear Models", *IDEI Working Paper*.

BOSTIC, R., G. STUART and G. PAINTER (2009): "Housing Wealth, Financial Wealth, and Consumption: New Evidence from Micro Data", *Regional Science and Urban Economics,*39 (1) pp. 79-89.

BRZOZOWSKI, M., M. GERVAIS, P. KLEIN and M. SUZUKI (2010): "Consumption, Income, and Wealth Inequality in Canada", *Review of Economic Dynamics*, 13 (1), pp. 52-75.

CARROLL, C., K. DYNAN and S. KRANE (2010): "Unemployment Risk and Precautionary Wealth: Evidence from Households' Balance Sheets", *The Review of Economics and Statistics*, 85:3, pp. 586-604.

CHEN, X., H. HONG and E. TAMER (2005): "Measurement Error with Auxiliary Data", *The Review of Economic Studies*, 72:2, pp. 343-66.

CHEN, X., O. LINTON and I. VAN KEILEGOM (2003): "Estimation of Semiparametric Models when the Criterion Function is Not Smooth", *Econometrica*, 71:5, pp. 1591-1608.

CHESHER, A. and L. NESHEIM (2006): "Review of the Literature on the Statistical Properties of Linked Datasets", *DTI Occasional Papers No 3*.

CROSS, P. and C. MANSKI (2002): "Regressions, Short and Long", *Econometrica*, 70(2), pp. 357-68.

DEVEREUX, P. and G. TRIPATHI (2009):"Optimally Combining Censored and Uncensored Datasets", *Journal of Econometrics*, 151(1) pp. 17-32.

D'ORAZIO, M., M. DiZIO AND M. SCANU (2006); *Statistical Matching, Theory and Practice*, Wiley.

DUGUNDJI, J. (1951): "An Extension of Tietze's Theorem", *Pacific Journal of Mathematics*, 1:3, 353-67.

FAN, Y. and S. PARK (2014): "Nonparametric Inference for Counterfactual Means: Bias-Correction, Confidence Sets and Weak IV", *Journal of Econometrics*, 178(1), pp. 45-56.

FAN, Y., R. SHERMAN and M. SHUM (2014): "Identifying Treatment Effects under Data Combination", *Econometrica*, 82(2), pp. 811-22.

FANG, H., M. KEANE and D. SILVERMAN (2008): "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market", *Journal of Political Economy*, 116(2) pp. 303-50.

FANG, Z. and A. SANTOS (2014): "Inference on Directionally Differentiable Functions", unpublished manuscript.

GOLDBERGER, A. (1991): *A Course in Econometrics*, Harvard University Press.

HAYASHI, F. (2001): *Econometrics*, Princeton University Press.

HELLERSTEIN, J. and G. IMBENS (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting", *The Review of Economics and Statistics*, 81:1, pp. 1-14.

HIRANO, K. and J. PORTER (2012): "Impossibility Results for Non Differentiable Functionals", *Econometrica*, 80:4 pp. 1769-90.

HIRIART-URRUTY, J. and C. LEMARECHAL (2004): *Fundamentals of Convex Analysis*, Springer.

HIRUWAKA, M. and A. PROKHOROV (2014): "Consistent Estimation of Linear Regression Models Using Matched Data", *Business Analytics Working Paper Series*, No WP201403.

ICHIMURA, H. and E. MARTINEZ-SANCHIS (2010): "Identification and Estimation of GMM Models by Combining Two Data Sets", *unpublished manuscript*.

INOUE, A. and G. SOLON (2010): "Two-Sample Instrumental Variables Estimators", *The Review of Economics and Statistics*, 92:3, pp. 557-61.
IPSOS MORI (2011): *Data Fusion - A White Paper by Ipsos MORI*, available at http://www.ipsos-mori.com/DownloadPublication/1425_IpsosMediaCT_WhitePaper_DataFusion_Jun2011.pdf
JAPELLI, T., J. PISCHKE and N. SOULELES (1998): "Testing for Liquidity Constraints in Euler Equations With Complementary Data Sources", *The Review of Economics and Statistics*, 80:2, pp. 251-62.
KLEVMARKEN, N. (1982): "Missing Variables and Two-Stages Least Squares Estimation from More than One Dataset", in *1981 Proceedings of the American Statistical Association*, pp. 156-61.
KOMAROVA, T., D. NEKIPELOV and E. YAKOVLEV (2012): "Identification, Data Combination and the Risk of Disclosure", *unpublished manuscript*.
MAGNAC, T. and E. MAURIN (2008): "Partial Identification in Binary Choice Models: Discrete Regressors and Interval Data", *Review of Economic Studies*, 75:4 pp. 835-64.
MEGHIR, C. and M. PALME (1999): "Assessing the Effect of Schooling on Earnings Using a Social Experiment", *The Institute for Fiscal Studies*, Working Paper No W99/10.
MOLINARI, F. and M. PESKI (2006): "Generalization of a Result on Regressions, Short and Long", *Econometric Theory*, 22(1) pp. 159-63.
MORIARITY, C. and F. SCHEUREN, (2003): "A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics*, 21:1, pp. 65-73.
PACINI, D. (2016): "A Confidence Interval for a Parameter Restricted by Nonsmooth Inequalities", *manuscript*.
POIRER, A. and N. ZIEBARTH (2014): "A Simple Estimator for Datasets with Non-Unique Identifiers", *Unpublished manuscript*.
RACHEV, S. and L. RUSCHENDORF (1998): *Mass Transportation Problems, Volume I: Theory*, Springer.
RASSLER, S. (2002): *Statistical Matching*, Springer.
RIDDER, G. and R. MOFFITT (2007): "The Econometrics of Data Combination", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Volume 6B, Elsevier.
ROCKAFELLAR, T. (1970): *Convex Analysis*, Princeton University Press.
RODGERS, W. (1984): "An Evaluation of Statistical Matching", *Journal of Business and Economic Statistics*, 2:1, 91-102.
RUBIN, D. (1986):"Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics* 4:1, pp. 87-94 .
SHAPIRO, (1990): "On Concepts of Directional Differentiable Functions", *Journal of Optimization Theory and Applications*, 66:3, pp. 477-87.
THE NIELSEN COMPANY (2009): *Introduction to Nielsen Data Fusion*, available at http://www.nielsen.com/content/dam/corporate/us/en/docs/solutions/Nielsen-Introduction-to-Data-Fusion.pdf
Van der VAART, A. (1998): *Asymptotic Statistics*, Cambridge University Press.
VITALE, R. (1979): "Regression with Given Marginals", *The Annals of Statistics*, 7(3), pp. 653-658.